Doing Good Badly?
Philosophical Issues Related to Effective Altruism

Michael Plant
St Cross College
University of Oxford
August 2019

Thesis for the degree of Doctor of Philosophy in Philosophy
Word count: 68,425

Abstract

Suppose you want to do as much good as possible. What should you do? According to members of the effective altruism movement—which has produced much of the thinking on this issue and counts several moral philosophers as its key protagonists—we should prioritise among the world's problems by assessing their scale, solvability, and neglectedness. Once we've done this, the three top priorities, not necessarily in this order, are (1) aiding the world's poorest people by providing life-saving medical treatments or alleviating poverty itself, (2) preventing global catastrophic risks, such as those posed by nuclear war or rogue artificial intelligence, and (3) ending factory farming.

These claims are both plausible and striking. If correct, they should prompt a stark revision of how we approach our altruistic activities. However, the project of determining *how* to do the most good—as opposed to say, *whether* we should do the most good—has only recently, within the last ten years, become the subject of serious academic attention. Many key claims have not yet been carefully scrutinised. This is a cause for concern: are effective altruists doing good badly?

In this thesis, I critique and develop some of the latest claims about how individuals can do the most good. I do this in three areas: the value of saving lives (preventing premature deaths), how best to improve lives (making people happier during their lives), and cause prioritisation methodology (frameworks for determining which problems are the highest priorities). In each case, I raise novel theoretical considerations that, when incorporated, change the analysis.

Roughly speaking, my main conclusions are (1) saving lives is not as straightforwardly good we tend to suppose, may not be good at all, and is not clearly a priority; (2) happiness can be measured through self-reports and, based on the self-reported evidence, treating mental health stands out as an overlooked problem that may be an even more cost-effective way to improve lives than alleviating poverty; (3) the cause prioritisation methodology proposed by effective altruists needs to be moderately reconceptualised and, when it is, it turns out it is not as illuminating a tool as we might have thought and hoped.

Acknowledgements

Although none of the following people were kind enough to write the thesis for me, I am nevertheless hugely grateful to them for their help in helping me to write it.

This thesis was primarily supervised by Hilary Greaves and Peter Singer and, for two periods when Hilary was on maternity leave, Frank Arzentius and William MacAskill. They have all contributed enormously to improving my work and thinking. It has been an honour to learn from them. I particularly want to thank Hilary for her determined (although usually in vain) attempts to improve the rigour of my work and Peter for supervising me even though, not being part of the Oxford Philosophy Department, he only had a moral, and not an institutional, obligation to do so.

For their comments on or help with various aspects of this work, I thank: Simon Beard, Joe Bowen, Tom Douglas, Tommy Francis, Paul Frijters, John Gustafsson, Alex Heape, Michelle Hutchinson, William James, Guy Kahane, Caspar Kaiser, Todd Karhu, James Kirkpatrick, Christian Krekel, Kacper Kowalcyzk, Richard Layard, Jeff McMahan, Michael Masny, Andreas Mogensen, Aidan Penn, Theron Pummer, Gary O'Brien, Karl Overdick, William Rooney, Korbinian Ruger, Julian Savulescu, Rhys Southan, members of the Effective Altruism community too numerous to name, and those I have carelessly forgotten but do not deserve to be forgotten. I am thankful to the audiences of the Applied Ethics Graduate Discussion Group, the Cumberland Lodge Colloquium on Population Ethics, Effective Altruism Global (and other, similar events), the Ockham Society, the International Society of Utilitarian Studies, the International Association for the Philosophy of Death and Dying, and those who mistakenly wandered into my lecture series *Philosophical Issues Related to Maximising Happiness* in Michaelmas term 2018, which previewed much of what follows.

I thank my parents, William and Penny, and brother, Edward, for their love and support. It has made my questionable choice of career possible. For many years, my parents, rightly concerned about my employment prospects, would ask, "Michael, what are you going to *do* with a degree in philosophy?" After nearly three philosophy degrees, I still do not have a good answer to that question (or any other, for that matter). I thank my friends, whose refusal to take either me or my research seriously

has kept me going. I thank my girlfriend, Hayley Capp, for her encouragement throughout; I can only apologise for the number of times I have insisted on talking at her when I was stuck on a problem.

Finally, I would like to thank Patrick Kaczmarek, who read and then discussed nearly every part of this thesis with me, typically over a pint and often in The Rusty Bicycle. All remaining errors are his.

Table of contents

Introduction

Suppose you want to do as much good as possible. What should you do? According to members of the effective altruism movement—which has produced much of the thinking on this issue and counts several moral philosophers as its key protagonists—we should prioritise among the world's problems by assessing their scale, solvability, and neglectedness. Once we've done this, the three top priorities, not necessarily in this order, are (1) aiding the world's poorest people by providing life-saving medical treatments or alleviating poverty itself, (2) preventing global catastrophic risks, such as those posed by nuclear war or rogue artificial intelligence, and (3) ending factory farming.[1]

These claims are both plausible and striking. If correct, they should prompt a stark revision of how we approach our altruistic activities. However, the project of determining *how* to do the most good—as opposed to say, *whether* we should do the most good—has only recently, within the last ten years, become the subject of serious academic attention. Many key claims have not yet been carefully scrutinised. This is a cause for concern: are effective altruists doing good badly?

In this thesis, I critique and develop some of the latest claims about how individuals can do the most good. I do this in three areas: the value of saving lives (preventing premature deaths), how best to improve lives (making people happier during their lives), and cause prioritisation methodology (frameworks for determining which

[1] (Singer, 2015) and (MacAskill, 2015) are the two original books advocating effective altruism. For a more recent articles setting out and defending effective altruism, see (MacAskill, 2018) and (Pummer and MacAskill, 2019).

problems are the highest priorities).[2] In each case, I raise novel theoretical considerations that, when incorporated, change the analysis.

Roughly speaking, my main conclusions are (1) saving lives is not as straightforwardly good we tend to suppose, may not be good at all, and is not clearly a priority; (2) happiness can be measured through self-reports and, based on the self-reported evidence, treating mental health stands out as an overlooked problem that may be an even more cost-effective way to improve lives than alleviating poverty; (3) the cause prioritisation methodology proposed by effective altruists needs to be moderately reconceptualised and, when it is, it turns out it is not as illuminating a tool as we might have thought and hoped.

Before I provide some context to the chapters and outline their contents, I make three remarks.

First, the aim of the thesis is to see what follows, given particular moral theories, rather than to evaluate which moral theory is correct. Philosophers have focused extensively on the latter over the years, which is why I turn to the former to make my contribution; we will shortly see it reveals a rich bounty of interesting theoretical questions. Hence, this thesis is of the flavour 'if X were true, then the surprising result is Y' without arguing for the truth of any particular X.

Second, I expect the subject matter of the thesis to be of greatest interest to those (such as this author) who believe that, if we want to do the most good, the practical priority is to make people happy, rather than make happy people, or prevent the

[2] I note my usage of 'improve lives' is arguably non-standard. Ordinarily, improving lives would refer to increasing individuals' well-being during their lives, rather than, less ecumenically, their happiness. As I focus specifically on happiness, where such specificity is required, as opposed to any other good that may constituent well-being (in whole or in part), the non-standard usage is more appropriate.

making of unhappy animals.[3] I focus on saving and improving lives and offer no argument here that those who already prioritise something besides these need to reconsider their position.

Third, the chapters have, for the most part, been written so each is comprehensible without having read the others; as I am not advocating a particular moral view or building towards a single conclusion, this seemed the easiest way to structure the thesis. The final two chapters are something of an exception to this—both develop ideas from earlier in the thesis and will be less understandable if they are read by themselves. Given this structure, this introductory chapter serves the important role of putting the different chapters into context and explaining how the arguments connect to one another.

There are seven chapters. Chapters 1 to 3 concern saving lives, chapters 4, 6, and 7 are about improving lives, chapters 5 and 6 relate to cause prioritisation. To explain the overlap, chapter 6 proposes a new approach to cause prioritisation and applies it in the case of improving lives. The context and contents of the chapters now follow.

Prima facie, saving a life does a great deal of good. Given the best estimates are that we can prevent a premature death for, on average, only a few thousand pounds if we donate to certain effective charities, it is easy to see why one would think saving lives is the most good you can do.[4] However, on closer inspection, matters are not so straightforward.

---

[3] This paraphrases (Narveson, 1973) who at p. 70 states "morality is in favour of making people happy but neutral about making happy people". I add a further reference to animals to accommodate the fact some think preventing the existence of unhappy animals is high priority.
[4] According to the analysis of charity evaluator (GiveWell, 2019a)

Chapter one raises one complication: what if, when considering the value of saving lives, we account for the fact that most people are meat eaters? Peter Singer has famously argued that eating meat is wrong on the grounds of the animal suffering this causes.[5] Singer has also argued that we would be morally required to jump into a shallow pond to save a drowning child even if this would ruin our expensive clothes.[6] I argue that, assuming meat eating is wrong on the grounds of the animal suffering caused, it is plausible that meat eaters cause so much suffering that saving the life of an average stranger is bad. If we grant the compelling principle that we are not required to bring about the worse of two outcomes (or slight variants of this principle), then it follows that we are not required to save lives, even in cases where we can easily rescue someone. I do not seek to argue meat eating is wrong on grounds on animal suffering; this argument investigates what follows if we accept that view. The more general result of the argument is that, even if we do not think accounting for meat eating is sufficient to make saving lives bad, it will reduce the value of doing so by some amount. I note this is the only chapter in which I engage in normative issues, that is, about what we ought to do; all the other chapters are concerned solely with the value of outcomes.

Chapter 2 adds another wrinkle to determining the value of saving lives. Many people seem to accept the 'Intuitive View', that saving lives is good and—as the Earth is overpopulated—that averting new lives is good too. Although it is not widely recognised, the Intuitive View is in internal tension—another way to reduce population size is by not saving lives. This raises some questions, for instance, whether and how the Intuitive view could be true. I develop Greaves' earlier analysis

---

[5] (Singer, 1975)
[6] (Singer, 1972)

of this topic and then investigate both how probable the Intuitive View is and what the practical implications are of accounting for under/overpopulation.[7] I do this by first assuming Totalism is the correct theory of population ethics, and second assuming a Person-Affecting View. I argue that the Intuitive View is distinctly unlikely on Totalism: it is more probable that one of saving lives or averting lives is bad. Further, were the Intuitive View true, the value of each of saving and averting lives would be small—far smaller than we would expect—and thus both are relatively unpromising interventions if the aim is to do the most good. This result is problematic for Peter Singer who seems to hold the Intuitive View, Totalism, and that life-saving and life-averting charities are among the most-effective giving opportunities—this combined position is highly improbable.[8] On a Person-Affecting view, the Intuitive View is (unsurprisingly) far more likely. If both the Person-Affecting and Intuitive Views are true, the values of saving and averting lives would be in large range (how large is specified in the chapter). The general result is that we don't know the values of saving or averting lives unless we know how under- or overpopulated the world is. I close by briefly arguing that it's not obvious where the world is in relation to optimum population size (whether we think just this generation or all generations matter); hence it's correspondingly unclear what the value of saving lives is on this wider analysis.

Chapters 1 and 2 consider some problems that arise when we account for the *other-regarding* impacts of saving a life—the impact saving lives has on others (human and non-human). Many people seem to think saving lives is the most good we can

---

[7] (Greaves, 2015)
[8] See the recommendations made by (The Life You Can Save, 2019) a charitable organisation founded by Singer and named after his book of the same name (Singer, 2009).

do—the particular effective altruist suggestion is to save the lives of children in the developing world with cheap, effective health interventions. Those making this suggestion seem to be basing this primarily on the *self-regarding* value of saving a life—the value associated to the person whose life is saved. Chapter 3 takes a step back and asks whether, if we ignore these other-regarding effects, saving lives could still the most good we can do. I set out four commonly-held, but not exhaustive, views of the value of creating and ending lives (such accounts are a combination of a population axiology with an account of the badness of death). In each case, I claim it's not obvious that saving lives is the most cost-effective option—either there is an alternative that seems similarly cost-effective or the view is sufficiently underdetermined that we cannot straightforwardly make the comparison.

Taken together, the conclusions of the first three chapters are that (1) accounting for the other-regarding effects of saving lives makes it far less clear that saving lives is good, or even if it is good, and (2) even if we ignore these other-regarding effects, saving lives is still not clearly a priority on any of a commonly-held range of views.

In chapter 4, the focus moves on to how best to improve lives, that is, to make people happier during their lives. Singer and MacAskill seem to suggest that the best way to do this is by alleviating global poverty.[9] While this suggestion is also, on its face, highly plausible, I argue this is not so clearly the case either.

Chapter 4, the longest in the thesis, starts with the observation that, while MacAskill and Singer seem to hold that well-being consists in happiness, they do not seem to make use of the 'subjective well-being' (SWB) literature from psychology and economics—where individuals provide self-reported measures of their moment-by-

[9] (Singer, 2015), (MacAskill, 2015)

moment happiness and/or their overall life satisfaction. Instead, MacAskill and Singer rely on more conventional metrics for well-being such as income and Quality-Adjusted Life-Years (QALYs), a measure of health. I suggest four possible objections to using self-reports, rather than of any other metric(s), to determine what increases happiness: (1) happiness is not measurable through self-reports; (2) individuals' scores are not interpersonally cardinally comparable (i.e. a one-point increase for one person on a 0-10 scale is equivalent to a one-point increase for anyone else); (3) there isn't enough available evidence of such self-reports to use them to guide decision-making; (4) it is unnecessary to use such data as they wouldn't change the priorities anyway. Chapter 4 addresses the first objection fully and the latter three partially. I argue that, despite the doubts, 'subjective well-being' (this umbrella term includes both happiness and life satisfaction) can be measured through self-reports. I explain why it's broadly reasonable to treat the self-reports as interpersonally cardinally comparable even though we cannot be certain of this. I also argue that even if the scales are not interpersonally cardinally comparable that should does not present a sufficient reason not to use them. Regarding the latter two objections, I provide suggestive evidence that there is enough data on SWB to inform our actions and it may lead us towards different priorities: the clearest case is mental health, which stands out as a major cause of unhappiness and has not been considered an important issue by effective altruists so far. A more compelling reply to the final two objections is offered in chapter 7.

As chapter 4 revealed there is a new, potential priority if we want to make people happier, this prompts the question of what method, in general, we should use to determine which of the world's problems are more important than others (more important in the sense that it allow us to do more good). MacAskill notes that typical

priority-setting method used by effective altruists is the three-factor 'cause prioritisation' framework, on which problems are evaluated by considering their scale, solvability, and neglectedness.[10] The obvious thought is we should just apply this method to the domain we are focusing on—improving lives. The problem is that, while the three-factor method seems to capture something important, no careful analysis has been done of the method and there are a number of open questions about how exactly it works. For instance, MacAskill implies it's possible to evaluate problems via the framework *prior* to making quantitative cost-effective assessments of the solution to those problems. This is somewhat mysterious—how it is possible to evaluate a problem(/cause) prior to assessing the cost-effectiveness of particular solutions(/interventions) to that problem?

In chapter 5, I set out the cause prioritisation method, clarify its workings, and address some outstanding questions. I suggest the priority-setting process should be partially reconceptualised. The most important conclusion is that using the scale-neglectedness-solvability framework to evaluate the cost-effectiveness of a problem (e.g. poverty) ultimately relies on how carefully we have estimated the cost-effective of particular solutions to that problem (e.g. particular poverty interventions). Hence, the cause prioritisation method turns not to offer much in the way of useful shortcuts for determining what the most important problems are: if we want to find out how to do the most good, we'll have to carefully assess the many different ways we might solve the problems in front of us.

Having identified the limits of the cause prioritisation methodology in chapter 5, chapter 6 proposes a new method, 'cause mapping', in order to help organise our

---

[10] (MacAskill, 2018)

search for potential solutions to the problems we are interested in. In essence, cause mapping involves breaking the priority-setting process into a series of distinct steps and considering the relevant possibilities at each step in order to help us structure our thinking. I then apply the cause mapping approach to the question of how best to improve lives. The result is a 'long list' of options which, *prima facie*, seem to be highly promising ways to improve lives. In addition to mental health and poverty, I suggest pain, positive education (i.e. teaching resilience and life skills in schools), and drug policy reform stand out. As such, I set some further possibilities that have not yet been seriously considered by any effective altruists. What is required next is a careful empirical investigation of these options.

While I am unable to assess all these possibilities, I do examine one in some detail in chapter 7. As Singer and MacAskill draw their charity recommendations from GiveWell, a charity evaluator, I attempt a first-pass cost-effectiveness analysis using SWB scores, which compares GiveWell's eight top recommendations to a developing world mental health organisation, StrongMinds. StrongMinds seems roughly on a par, in terms of cost-effectiveness, to GiveWell's top-rated life-improving recommendations. Comparing life-saving to life-improving charities turns out to be highly sensitive to an unresolved methodological question (where on a 0-10 life satisfaction scale is the 'neutral' point that is equivalent to non-existence) and I am unable to draw conclusions here. The analysis in this chapter also allows me to decisively meet the latter two objections raised in chapter 4 about the use of SWB data, by showing there is enough available evidence SWB data to guide decision-making and that it does indicate new priorities.

Have effective altruists been doing good badly? I do not go so far as to claim *that*—not least because it's unclear how we would evaluate this question. However, in this

thesis, I am able to set out a range of important, novel theoretical considerations. This challenges current thinking about how to the most good in several key areas and shows how those looking to do the most good can do good even better.

Chapter 1: The Meat Eater Problem

0. Abstract

I argue, *pace* Singer, that if we're wrong to eat meat because of the suffering this causes then we would not be required to save an easily rescuable child from drowning in a shallow pond. At least, this follows if we grant certain other plausible normative and empirical assumptions.

1. Introduction

Consider the following familiar case:

> *Shallow Pond*: you are walking past a shallow pond and see a drowning child. You can easily rescue the child but doing so will ruin your new suit.[1]

What should you do? Nearly everyone accepts we are morally required to jump in. As Peter Singer explains: "this will mean getting my clothes muddy, but this is insignificant, while the death of the child would presumably be a very bad thing."[2] More generally, we accept the *Principle of Easy Rescue*:

> *The Principle of Easy Rescue*: we are required to save lives in *rescue cases*, one-off instances where we can physically save an average stranger at a trivial cost to ourselves.

[1] This case comes from (Singer, 1972)
[2] Ibid, p231

A second question: is it wrong to be a *meat eater*, someone who regularly consumes animal products produced from factory farms?[3] Many think it is and accept the *Weak Carnism Thesis*:[4]

> *Weak Carnism Thesis*: it's wrong to be a meat eater because of the animal suffering this causes.

Such people conclude we ought to become vegans or vegetarians. While there are other reasons one might believe meat eating is wrong—the environmental impact, the violation of animals' rights, etc.—I will only focus on animal suffering here as it is the easiest way to generate the problem I now raise.

Many believe both the Principle of Easy Rescue and the Weak Carnism Thesis are true. The most obvious example is Singer: he is the originator of the *Shallow Pond* case and encourages people to give to life-saving charities;[5] he has publicly campaigned against factory farming on animal suffering grounds and advocated veganism for many years.[6] The conjunction of the two beliefs is widely held among moral philosophers and, increasingly, in society at large.

This essay argues the beliefs are incompatible, given additional plausible empirical and normative assumptions and, faced with a choice of which to give up, it is the Principle of Easy Rescue that should be abandoned. This conclusion is reached in five moves.

---

[3] I will use the term 'meat eaters' as a shorthand for 'factory-farm-produced-animal-product-consumers'. This is revisionary: someone who consumes battery-farmed eggs, but no animal flesh would be a meat eater; someone who ate meat only from free-range animals would not be. Given the state of the world, this revision is unproblematic (see footnote 18).

[4] For instance, (Singer, 1975)

[5] (Singer, 1972)

[6] (Singer, 1975)

First, I observe a widely-held intuition is that we are *not* required to save lives, even if we can do so easily, when the consequences would be bad *enough*. Consider:

> *Drowning Dictator*: you live in a country ruled by a dictator. One day, when walking past a pond you see him drowning—he has a distinctive appearance, and you recognise him from his ubiquitous propaganda posters. You can easily rescue him, but you realise that doing so will not only ruin your expensive new shoes, but he will go on to torture and terrorise thousands of people in the future.

Presumably, few think we are *required* to save the dictator. When, exactly, are the consequences bad *enough*? Consider:

> *The No Requirement to cause Acts of Greater Evil* ('NRAGE') *Principle*: you are not required to prevent something bad from happening if, in so doing, this will bring about an even worse outcome.

How would NRAGE apply to *Drowning Dictator*? We have a choice between (A) allowing a bad thing or (B) preventing that bad thing, but in so doing bringing about something even worse. NRAGE holds we are *not* required to do (B). The principle is minimal: it still leaves open whether both choices are permitted or that the better outcome is mandatory. Further, NRAGE does not specify how the value of outcomes is assessed, only that, however this is done, we are not required to pick the worse outcome.

While the wording of NRAGE will be unfamiliar, its plausibility is straightforward. Consequentialists hold you are *required* to choose the better option; hence you are *not* required to choose the worse one. Many non-consequentialists accept there is a *lesser-evil justification* for doing harm: we are permitted to cause harm if this will prevent a *substantially* greater harm. The justification for the harm needing to be

substantially greater is the familiar doing/allowing distinction: the normative badness of doing harm is such that a substantially larger harm needs to be prevented to overcome this normative badness.[7] Importantly, however, *Drowning Dictator* is not a case of causing a harm to prevent a greater evil; rather, it seems akin to a choice between allowing a lesser evil—the dictator's death—and allowing another, greater evil—the suffering that results if the dictator lives. In a choice between allowing a lesser and a greater evil, it seems that we are not required to allow the greater evil. Non-consequentialists might think we are permitted to allow either evil, or we are required to pick the lesser of the two, but it seems counter-intuitive to think we could be required to pick the greater evil (I will return to this later).

Second, I argue that, if the Weak Carnism Thesis is true, then the *Strong Carnism Thesis* is plausible:

> *Strong Carnism Thesis*: meat eaters cause a sufficiently large amount of animal suffering via their diet that the consequences of saving a stranger's life in rescue cases are (in expectation) worse than those of not saving them.

This is mainly an empirical claim. I argue that meat eaters cause so much suffering that saving their lives in rescue cases would be worse than not saving them. At least in a country like America, given around 95% of people are meat eaters and 99% of meat comes from factory farms, it is not difficult to infer that, if saving *meat eaters* is bad, saving a randomly sampled *stranger* is bad in expectation too.[8]

---

[7] E.g. see (Frowe, 2018) for a discussion of the lesser-evil justification.

[8] According to the (National Chicken Council, 2018) less than 1% of US chickens are free range. (Sentience Insitute, 2018) estimate that 99.5% of land animals – meat chickens, egg chickens, turkeys, pigs and cows – in the US are reared in factory farms. Note the vast majority of animals are chickens: Americans eat 8 billion chickens for meat each year, but only 90 million cows and 65 million pigs. A consumer report by (GlobalData, 2017) found 6% of American self-identify as vegans. A poll conducted by (Vegetarian Resource Group, 2016) of 2,000 adult Americans found 3.4% ate a solely vegetarian diet, i.e. they agreed with the statement "I never eat meat, fish, seafood, or poultry".

Third, I observe it is inconsistent to believe the Principle of Easy Rescue, NRAGE, and the Strong Carnism Thesis. The Principle of Easy Rescue states we are *required* to save strangers (at least, where we can do so at minimal cost to ourselves). NRAGE and the Strong Carnism Thesis together entail that although saving the life of a random stranger would prevent something bad (i.e. the stranger's death), we are *not required* to prevent this because saving the random stranger would be worse than not saving them. These three principles are jointly inconsistent: they require us to both save and not save strangers in rescue cases. To avoid inconsistency, we must revise at least one of the three.

Fourth, I assume that, faced with this inconsistency, most are tempted to conclude NRAGE is false and the Principle of Easy Rescue true. I show this strategy is less promising than it looks. While we can revise NRAGE, for the two revisions I consider we can also make a corresponding adjustment to the Strong Carnism Thesis and re-generate the inconsistency.

Fifth, now that we are pushed to abandon the Principle of Easy Rescue, I explain why this conclusion is less implausible, on reflection, than it appears.[9]

The result is that those who believe it is wrong to eat meat because of the suffering this causes to animals should find it plausible that meat eaters cause enough

---

[9] Since writing up this argument, I discovered the idea that concerns for meat eating could make saving humans lives bad has been discussed, but seemingly not written about, in the effective altruism and animal advocacy communities. One written treatment I found is (Weathers, 2016), who considers whether the idea eating meat should cause some reduction in the value of saving lives. Weathers concludes meat eating may not be bad (the created animals might be happy) and, even if it is, saving lives is still good. I take my contribution, then, to be (a) rigorously examining (in writing) what we might need to hold to make saving lives bad (in light of concerns about meat eating), (b) assessing whether those things do hold, (c) proposing a plausible a normative principle, NRAGE, which is inconsistent with the Strong Carnism Thesis and Principle of Easy Rescue, and thus (d) showing the meat eater problem is not only an awkward issue for consequentialists.

suffering (on different specifications of 'enough') that we are not required to save lives, even when we can do so easily.

Then, having stated the argument, I set out two further (troubling) implications. First, those, such as Singer, who hold morality requires us to choose the better of two options (either *tout court* or when it is not too demanding), will hold it is *wrong* to save the lives of strangers.[10] Second, accounting for humans' impact on animals may substantially alter our charitable priorities.

This essay is structured as follows. Section 2 argues those who accept the Weak Carnism Thesis should find the Strong Carnism Thesis plausible. Section 3 considers how we might tinker with NRAGE. Section 4 argues giving up the Principle of Easy Rescue is the correct response to the inconsistency. Section 5 discussions two implications of the argument. Section 6 concludes.

2. If the Weak Carnism Thesis is true, how plausible is the Strong Carnist Thesis?

In this section, I ask how plausible the Strong Carnism Thesis is given the Weak Carnism Thesis. The purpose of this chapter is not to persuade those who find the Weak Carnism Thesis implausible, but to explore what might follow if this thesis is true. While I do not claim the Strong Carnism Thesis is clearly true, there is a credible case for it, and it is not something that can reasonably be rejected out of hand.

---

[10] (Frowe, 2018) gives a non-consequentialist argument for the view that we are required to do the lesser evil where this is not overly demanding.

Before we go any further, let me state the four sub-claims I assume which, together, entail the Weak Carnism Thesis. First, animals in factory farms lead lives with net negative well-being. Second, the individual purchasing decisions of meat eaters cause animals to be created.[11] Third, creating unhappy lives, lives with net negative lifetime well-being, is bad.[12] Fourth, the total negative well-being caused by creating unhappy animals is sufficiently large, relative to the benefit a meat eater gets from eating meat (compared to the scenario in which they do not) to make being a meat eater wrong.[13]

I evaluate the Strong Carnism Thesis in three steps. First, I ask whether the *self-regarding value* of saving meat eater's lives–that is, the value associated with the individual whose life is saved—is greater than the *animal impact* of that life—that is, the net *dis*value the person causes to animals through their diet. Animal impact is one part of the *other-regarding impact* of saving a life, that is, the impact saving a life has on everyone else. Second, I consider the other other-regarding impacts. Third, I account for the fact not all humans are meat eaters.

2.1 Comparing self-regarding value of meat eaters' lives to their animal impact

By value(/disvalue), I will have happiness(/unhappiness) in mind, unless otherwise specified, where happiness is a net positive balance of pleasure over displeasure (unhappiness the reverse). I later discuss whether valuing goods besides happiness changes the result. For the moment then, the first question is effectively: do meat

[11] This is probably the most contested assumption. (Kagan, 2011) argues you do sometimes make a difference.
[12] Of the assumptions, this is the one I am most sympathetic to thinking is untrue.
[13] For the purposes of the argument, it's unimportant what would count as 'sufficiently large'.

eater cause more unhappiness to animals through their diets than they experience happiness themselves?

Suppose we're assessing the amounts of happiness experienced and unhappiness caused over the period of a year. We start by asking: how many years of animal life do meat eaters cause to be created for each year they eat meat? Meat eaters might eat lots of animals, but to keep things simple, let's assume meat eaters just eat chicken. This isn't a very distortionary assumption given the empirical data on animal consumption.[14] Matheny and Chan estimate that an average American would need to eat 82.6 chickens a year if they got all 20kg of their annual recommended amount of protein from chicken meat.[15] As chickens live around 7 weeks before being killed, meat eaters will, therefore, create 10.8 chicken life-years each year they live. To simplify the later calculations, I round this number down to 10.

Given that the Weak Carnism Thesis is true, this means a chicken's life is, on average, unhappy. I assume that human life is, on average, happy. Next, we ask: how happy are the humans compared to the chickens they create? Using this 10-to-1 number, if the humans are less than ten times happier than the chickens are *un*happy, then, all else being equal, it is *worse* to save the life of an average meat eater than not save them. As a shorthand, I'm going to use "*HA*" to refer to the average happiness level of humans divided by the average happiness of the factory-farmed animals. To reduce confusion, I'll use the absolute value of *HA* (the

---

[14] (Matheny and Chan, 2005) point out that broiler chickens constitute 60% of all farm animal life-years in the US, broiler breeders 3.5%, egg-laying chickens 19.6%, which totals 83.1%. As broilers and egg-laying chickens produce roughly the same amount of protein per life-year, 1.8kg and 1.6kg, respectively, to get 20kg of protein from just chickens would require 10.8 broiler chickens or 12.5 layer chickens.

[15] Ibid. p585.

magnitude of the number without regard to its sign). Thus, supposing humans are at happiness level 1 and the animals at -1, then *HA* would be 1.

This second question is harder to answer. We might attempt to do so by asking ourselves "are humans' capacities for happiness more than 10 times greater than those of chickens?" We might doubt chickens can feel great happiness—they will never experience J S Mill's 'higher pleasures' or Parfit's 'the best things in life', whatever these are—and conclude meat eaters are at least 10 times happier than the average chicken is unhappy.[16] However, capacities are not necessarily relevant: we need to know the *average* magnitudes of happiness and unhappiness that are experienced by humans and chickens, not what their *capacities* are.

To think about average magnitudes, we should instead ask the following question, assuming we are as happy as the average meat eater: "how happy am I during an average hour of my life, how unhappy do I think a factory-farmed chicken is during an average hour of its life, and what is the relative difference of these two states?" That may still seem hard to answer. So, here are a couple of alternative formulations: "assuming how I feel *right now* is average, do I think I'm happier than an average factory-farmed chicken is unhappy right now, and if so, by what proportion?" and, alternatively, "how many chickens, living in a factory farm for an hour, would it take to experience the same amount of unhappiness as I experience happiness in an average hour of my life?"

While I sometimes experience great elation, the majority of my waking hours only feel mildly good. This isn't because I think I'm suffering from some problem, either mental or physical, it is just that my daily life doesn't come with very strong

---

[16] (Mill, 1861) Ch2; (Parfit, 1986)

emotions. Given how much time most of us spend working, how we feel during an average working moment is not far off the mean average happiness across our whole lives.

By contrast, the average moment a chicken spends in a factory farm is believably *at least as bad* as I feel good now. From the footage I have seen of chickens in factory farms, that experience seems to be one of cacophonous chaos: animals packed into tiny spaces, pushing past each other to acquire food.[17] The descriptions of life as a broiler are arresting:

> Chickens have been bred to grow at grossly accelerated rates, causing a number of skeletal and cardiovascular problems. At the ends of their lives, they live at densities of around a square foot per bird, and 90% cannot walk properly, due to skeletal disorders.[18]

Quoting Singer's writing:

> Chickens, reared in sheds that hold 20,000 birds, now are bred to grow so fast that most of them develop leg problems because their immature bones cannot bear the weight of their bodies. Professor John Webster of the University of Bristol's School of Veterinary Science said: "Broilers are the only livestock that are in chronic pain for the last 20 per cent of their lives. They don't move around, not because they are overstocked, but because it hurts their joints so much"[19]

---

[17] E.g. (Mercy for Animals, 2011). I assume, for advocacy purposes, they have shown the worst conditions. For a video produced by the agricultural industry, see (US Poultry, 2014).
[18] (Matheny and Chan, 2005) p582
[19] (Singer, 2006). The quote from Prof. Webster is from (The Guardian. 1991).

Regarding this last period of life, it is difficult to believe I experience more happiness than another creature experiencing chronic pain—I do not experience chronic pleasure.

Intuitions will differ, but *HA* is plausibly around 1. Given the earlier number that one year of human life creates around 10 years of chicken life, this means meat eaters are causing around ten times more unhappiness to animals than they experience happiness themselves. Note, we have reached this conclusion not by claiming chickens have terrible, torturous lives and humans have wonderful, elated ones, but instead by thinking that much of human life is only mildly happy, whereas life in factory farms seems to be at best stressful, and for considerable periods, painful.

I consider four objections.

First, quantifying average human happiness and comparing it to average animal unhappiness is too nonsensically speculative to even attempt.

This objection is not available to those who hold the Weak Carnism Thesis. Concluding that this is true requires, at least implicitly, quantifying the happiness that individuals get from eating meat, comparing that to how much those animals suffer, and then concluding that imposing such suffering is impermissible. Hence, the conjunction of the Weak Carnism Thesis with this objection would be motivationally unstable.

An additional problem with the objection is that if humans' and non-human animals' levels of happiness are practically incomparable, it is likely the Principle of Easy Rescue would be abandoned anyway. Consider:

Incomparability Thesis: if two options are incomparable in value, both are permissible.

Suppose once accepts this plausible thesis. It, combined with the Weak Carnism Thesis, entails that we are permitted to either save or not save the stranger in the *Shallow Pond*. Yet the Principle of Easy Rescue holds you are required to do so. Inconsistency strikes.

Second, we should be very uncertain about what *HA* is. Although I've listed this as an objection, it is not specific to this problem and has an uncontroversial solution: we should proceed as we do in other cases of uncertainty, by making a probabilistic estimate. We are quite happy to think that the average score of a fair six-sided die is 3.5.

Third, we could object that the meat eaters cause less animal impact, relative to the individual value of their lives, than I've claimed. There seem to be four different ways of pressing this, which I raise and address in turn. I call these *discounting* objections.

First, someone could claim that, when they think about it, *HA* seems to be a lot larger than I supposed.

I have nothing further to say on what *HA* seems to be. What I will say is the following. For the objection to work, i.e. to show that the meat eater experiences more happiness than they cause unhappiness via their diet, it can't be just that *HA* is a *bit* higher than 1 (assuming all else is equal). If the meat eater creates 10 chicken-years for each 1 of their own, *HA* needs to be fully 10 times higher and *that*, for the reasons given earlier, seems implausible.

The second discounting objection is to point out that the only value I've accounted for is happiness, that hedonistic welfarism (the view that well-being consists in happiness and well-being is the only thing of intrinsic value) is false and, when we account for these non-happiness and/or non-well-being values, this increases the individual value of the meat eater's life. There are different ways to be a non-hedonistic welfarist (i.e. one who rejects hedonistic welfarism). One could endorse the objective list as the correct account of well-being and claim human lives can contain things on this list, such as friendship or knowledge, which animals' lives cannot. Or, you could be a perfectionist, that is, hold that certain goods, possibly art and scientific achievement, can be valuable even if they have no impact on well-being.[20] Perfectionists could also then claim humans can produce such goods while non-human animals cannot.[21]

We can pose the same reply as before: are these discounts believably big enough to make a difference? Suppose the objective list is true and, for concreteness, well-being consists only in happiness, knowledge, friendship, autonomy and health. To make it better to save the meat eater, assuming HA is 1, someone would need to think something like the following: "the average person's life contains some happiness, knowledge, friendship, autonomy and health and these all contribute to their well-being. The contribution of non-happiness components is 9 times larger than that of the happiness components." Remember, we're considering an average

[20] (Parfit, 1986) p161 discusses the 'best things in life'.

[21] I omit discussion of preference satisfactionism, the view that well-being consists in one's preferences being satisfied, in the sense that what one prefers to be the case, is the case. This is both because it's unclear how to weigh the preference of the meat eater to live against the preferences of the animals not to be brought into a bad existence and because the view seems implausible.

person here, not a genius scientist, someone with lots of friends, exceptional health, etc. Consider the following case:

> *Happiness or other goods*: you have two options. Either (A), 9 people have their happiness reduced to the neutral-point, where they feel neither happy or unhappy or (B), 1 person has their knowledge, friendship, autonomy and health reduced to their respective neutral points (whatever these are).

If we think the non-happiness items on the objective list contribute nine times more well-being to someone's life than their happiness does, we would be indifferent between A and B. However, it is hard to believe these non-happiness components could be of such relative importance. This remains the case even if we add additional non-happiness items to the objective list.[22]

The third version would be to claim we should apply a *pure species discount*, that is, hold it is more valuable to give the same well-being increase to humans than to non-human animals.

We can make the same reply: if HA is 1 then, unless the chickens' well-being is discounted by at least 90%, we would still conclude it would be better not to save the human. I'm unaware of any living philosophers advocating such a large

[22] We can set up an equivalent case for perfectionist goods: if we think happiness and say, scientific knowledge, both have intrinsic value, to change the result we'd need to put an implausibly large value on the per-person value of scientific knowledge.

discount.[23] Shelly Kagan, who thinks such a discount *might* be justified, supposes it would only be small.[24]

A further problem for a steep pure species discount is that, however, the discount is justified, it seems likely to generate unacceptable results in 'marginal cases'. Suppose the discount is based on the superior rationality of humans. Very young children won't have this, which implies their suffering is unimportant. We might avoid this by saying that the capacity for rationality is what matters, not whether one has it at present. However, it now follows that humans with serious cognitive disorders—and thus who will never develop these rational capacities—should have their suffering severely discounted, even though their abilities to experience, pain, would be just as strong as those of other humans; this is counter-intuitive. Finding a suitable rationale for this discount is not straightforward.

The fourth and final discount is based on the idea that meat eaters will soon switch to 'clean meat', where animal cells are grown in a laboratory. Assuming no animals suffer in this process then, *prima facie,* the advent of clean meat will end the unhappiness that meat eating causes. Further, if the transition to clean meat—the 'transition' for short—happens soon, the Strong Carnism Thesis will certainly be false as the individual value of meat eaters' lives will be greater than their (negative) animal impact.

---

[23] (Kant, 1798) seemed to think we should give no weight to the interests of animals at all: "[the human being] is a being altogether different in rank and dignity from things, such as irrational animals, with which one may deal and dispose at one's discretion." Leaving aside how implausible this is, the target of this essay is not those—such as Kant—who would reject the Weak Carnism Thesis anyway.

[24] In print, (Kagan, 2016) claims a discount might—and also might not—be justifiable does not offer an account of big it might be. At his 2016 Uehiro lectures at the University of Oxford, Kagan suggested the discount would only be very small.

I raise three issues for this discount. First, the transition won't certainly happen and, if it did, wouldn't remove all meat eater-caused animal suffering: it might not be technologically possible to make meat cheap enough that consumers switch; presumably, not all consumers will switch even if it were possible; it will only apply to some products, e.g. one can already buy soy and rice milk, but people consume cow milk.

Second, even if we charitably assume the transition removes all the animal suffering that meat eaters cause, it is hard to believe it could happen soon enough. Suppose, for simplicity, the transition happens overnight and everyone switches. Let's say that Tim is the average stranger we're considering saving. Tim is median age, 30, and might expect to live about 40 more years. If *HA* is 10, then in 4 years, Tim will have caused as much animal unhappiness and he will experience happiness for the rest of his life. Hence, the transition would need to happen within 4 years, which is implausibly fast.[25]

Third, notice that the peculiarity of whether we should save strangers or not turns on how quickly a previously seemingly irrelevant technological development occurs. Relatedly, even if we thought the transition would occur quickly enough in the future, there is still the odd result that saving meat eaters lives in the past few year or decades *would have been bad*.

To summarise, taken individually, the discounting objections would need to be implausibly steep to make it better to save the meat eater.

---

[25] *Shallow Pond* supposes we're saving a child. Suppose this child is 10. For the same reasons as above, the transition would need to happen in 8 years to make it better to save the child. Again, this seems implausibly near.

However, the discounting objections could be pressed again: we should apply *multiple* discounts but apply each *modestly*. Perhaps the pure species discount is 50% rather than 90%, *HA* is 3 rather than 1, we give some weight to goods besides happiness, and the transition is 20 years away. Now it's *slightly* better to save the meat eater.

To be clear, while applying multiple discounts is not *ad hoc*, it would be *ad hoc* to adjust the strengths of the various discount solely in order to avoid the Strong Carnism Thesis. I concede that some will accept that Weak Carnism Thesis is true but, after giving careful consideration of the facts, including the discounts, conclude that the Strong Carnism Thesis is false. I did not claim I could prove the Strong Carnism Thesis beyond reasonable doubt. There is not space to discuss whether the Strong Carnism Thesis holds on various different combinations of discounts; I leave the interested reader to conduct their own analysis.

A final point on this objection. Only some philosophers will be able to apply all four discounts, given their preferred theoretical machinery. For instance, Peter Singer has argued against a pure species discount and is (now) a classical utilitarian (i.e. the right action is the one that maximises the sum total of happiness).[26] Therefore, the only way Singer could avoid the conclusion would be by claiming (a) *HA* is much higher than I supposed, (b) the transition will occur soon or (c) some combination of (a) and (b). For concreteness, if the transition occurred in 20 years in the manner stated, then *HA* would need to be 5, i.e. humans are 5 times happier than chickens are unhappy, to make it better to save Tim (on the analysis so far).[27] Given his

[26] Pertaining to discounts, see (Singer, 1975).
[27] If the transition happens in 20 years, Tim would create 200 chicken life-years and live for 40 more years himself.

writing on the suffering caused in meat production, I presume Singer would find it hard to believe humans are *that* much happier.

2.2 Factoring in the other parts of other-regarding impact

Animal impact is one of the other-regarding impacts of saving a life. The three others are: (1) the grief caused to friends and family from a bereavement, (2) on wider human society, (3) on wild animals. One might argue that, when we account for these impacts, it becomes more valuable to save lives—if they make it less valuable, that supports the Strong Carnism Thesis.

(1) is likely to be small, relative to the individual value of the life; I doubt many think the former is even 10% of the size of the latter. We should we consider (1) counterfactually and when we do, it will be smaller still: people are generally sad whenever someone dies, hence the disvalue of grief is the difference between the badness now and the badness later. Therefore, (1) seems to be a relatively minor consideration.

(2) and (3) could be substantial considerations but it is not clear that they are, or in which direction they point. Regarding (2), many think overpopulation is a serious issue, but it is not obvious that we are overpopulated, a point I will return to in chapter 2.7. Regarding (3), some think extra humans destroy nature and this is bad—as nature is valuable intrinsically and animals are happy—others think this is good—there is net suffering in nature, but this view is controversial.[28] These other-regarding impacts are too complicated to discuss here. As such, I am forced to set them to one side and proceed on the basis of a restricted analysis. I note, as I did

[28] E.g. (Tomasik, 2015) argues there is net suffering in the wild.

with clean meat, it would be odd if the Strong Carnism Thesis were false only once we had accounted for these facts, facts we do not normally consider when assessing the value of saving lives.

2.3 Adjusting the fact not all people are meat eaters

This analysis has considered whether saving a random *meat eater* would be bad. The Strong Carnism Thesis concerns whether meat eaters have a sufficiently large negative impact via their diet that the value of saving the life of an *average* person is bad in expectation. Adjusting the analysis to refer to a randomly-selected person, rather than meat eater, is unlikely to make much difference, at least in the developed world. As noted, 95% of people in America are meat eaters and 99% of animal products come from factory farms.[29] Hence this too is a minor consideration and unlikely to be decisive for the Strong Carnism Thesis.

To summarise the analysis in section 2 as a whole, if the Weak Carnism Thesis is true then the Strong Carnism Thesis is certainly plausible, although not obviously true.

3. Is NRAGE the problem?

Let's now suppose the Strong Carnism Thesis is true. The Strong Carnism Thesis, NRAGE and the Principle of Easy Rescue are inconsistent, so one or both of the latter two must be reformulated. For many—consequentialists and some non-consequentialists—NRAGE is impossible to deny. This section is addressed at those non-consequentialists who wish to reject NRAGE (presumably in order to hold onto the Principle of Easy Rescue). Recall:

---

[29] See footnote 18.

> *The No Requirement to cause Acts of Greater Evil* ('NRAGE') *Principle*: you are not required to prevent something bad from happening if, in so doing, this will bring about an even worse outcome.

I argue that abandoning NRAGE does not straightforwardly rescue the Principle of Easy Rescue. I discuss the two most promising ways to modify NRAGE and show these alternatives, when supplemented with a modified Strong Carnism thesis, regenerate an inconsistency with the Principle of Easy Rescue.[30]

One way we might be tempted to give up on NRAGE is by drawing a distinction between the *direct* and *indirect* impacts of our actions. Direct impacts are those we personally cause on someone, whereas indirect impacts, and those that occur through the intervening agency of an agent, we directly impact. Consider:

> *Surgeon*: we can either save 5 people with a drug or use that drug to save a surgeon who will save 100 different people.[31]

Kamm argues that "we should not ignore the 100 [whom the surgeon] could save. But we should also not give them their *full weight* as 100 individuals versus the 5 who need our resource to live" (emphasis added).[32]

The application here is that we directly save the stranger, but the suffering they cause to animals is an indirect impact, which has less weight. This makes it *relatively* easier to say we are required to save the stranger.

---

[30] Here are two ways to reject NRAGE I do not discuss: appealing to retributive justice or liability-based justifications for harm, on which it can be right, respectively, to punish or harm someone even if this produces worse consequences. Arguably, meat eaters are liable for their harm to animals and deserve to be punished. For discussion of liability, see (McMahan, 2002) and of retributive justice see (Walen, 2016)

[31] (Kamm, 1998)

[32] (Kamm, 1998) p108

However, appealing to the direct/indirect distinction is unlikely to make much difference. Even Kamm seems to think we should discount indirectly impact only slightly ("not give them their full weight"). We could thus qualify NRAGE by adding "once the appropriate discount for indirect effects has been applied". For concreteness, say we weight indirect impacts 10% less. To regenerate the inconsistency, we would only need to believe a slightly stronger version of the Strong Carnism Thesis on which the other-regarding disvalue of saving the stranger's life is only 10% greater than the individual value.

Second, one could reject NRAGE by rejecting *unrestricted aggregation*, the idea that many *small* goods(/bads) can ever be morally equivalent to a *substantial* good(/bad). Quoting Voorhoeve:

> Many believe both that we ought to save a large number from being permanently bedridden rather than save one from death and that we ought to save one from death rather than save a multitude of people (who would be well off in any case) from very minor harms, no matter how large this multitude.[33]

In Voorhoeve's terminology, we should aggregate only 'relevant' claims. Advocates of restricted aggregation, such as Voorhoeve, will not accept NRAGE as there will be cases when you are *required* to bring about the worse set of consequences (impartially viewed). One might claim that we should save one human life over *any number* of suffering chicken lives—the harms caused to the latter are not relevantly large enough to be aggregated against the loss to the former.

---

[33] (Voorhoeve, 2014) p64.

To evaluate this objection, we need two pieces of information. First, a method which tells us when claims are and are not relevant. Second, a clearer sense of the relative size of the human's and the chickens' claims. Let's look at these in turn. Voorhoeve proposes:

> [a] person's weaker claim is irrelevant in the face of another's stronger competing claim just in case, in a one-to-one comparison of these claims, the stronger claim ought to take priority from every person's point of view[34]

Common-sense morality judges it is permissible for you to let the stranger suffer, even when the stranger would suffer a smaller harm than you would have received, but only up to a point:

> If you face a very minor harm (such as an illness that will leave you bedridden for a day) and can either prevent this harm to yourself or prevent the death of a stranger, then it holds that you must save the stranger.[35]

Voorhoeve's test is that a claim is relevant compared to another when someone is permitted to prevent the smaller harm from happening to themselves instead of preventing the larger harm befalling another.

We turn to the second piece of information. The problem, as discussed in section 2.1, is that individuals will differ on the relative value of a human life compared to a chicken life. To find the upper limit of the chicken's value, we could take *HA* as 1, count happiness as the only value and deny a pure species discount. In this case, the life of a chicken consists of 7 weeks of unhappiness of the same magnitude as the average human experiences net happiness over 7 weeks.

---

[34] (Voorhoeve, 2014)
[35] Ibid p71.

What could the lower limit be? For our purposes, the lower limit would be the least amount of disvalue that the average chicken's life could have, relative to the meat eater's, *given the Strong Carnism Thesis is true*. If someone thinks the Strong Carnism Thesis is false, then they will, presumably, think they *are required* to save the stranger in the *Shallow Pond* anyway. In 2.2, I suggested that if the transition to clean meat would come in 20 years, then *HA* could be around 5. This would mean the 7-week life of a chicken has the same total of unhappiness as the average human experiences net happiness in 10 days.[36]

Next, we pose the following question: if you were facing a threat of having all your happiness removed for those periods, taking you to a neutral hedonic level for either, (a) 7 weeks or (b) 10 days, and you could either prevent this harm to yourself or prevent the death of a stranger, would you be permitted to let the stranger die? If we find this thought experiment unintuitive, we might ask: if faced with the prospect of being put into a coma for 7 weeks(/10 days), could we permissibly prevent the coma-experience from happening to ourselves rather than prevent the death of a stranger?[37]

Intuitively, it would be clearly permissible in the case of (a) and borderline in the case of (b). By comparison, Voorhoeve's example of a borderline case is the choice between losing your finger and saving a stranger. He adds: "[m]oreover, if your loss were moderately greater (several fingers, say) it would be quite clear that you could permissibly save yourself".[38] It strikes me that the severance of one finger and losing all your net happiness from 10 days of life are on a par: having one less finger

---

[36] 49 (days)/5 is 9.8, which I rounded up to 10.
[37] These analogies are rough: if you were in a neutral state for 2 weeks you could carry on with your normal life. You could not do this if you were in a coma.
[38] Ibid p81

wouldn't stop you doing anything, the pain would be temporary, and you might dine out on the story for years to come. If we appeal to common-sense morality, it seems more likely it is permissible rather than impermissible that one could prevent oneself being in a coma for 10 days at the cost of someone else's life. If this is correct, then the chickens' lives are relevant compared to the human's life even at the lower bound and, therefore, we can aggregate the former against the latter. Assuming the Strong Carnism Thesis is true, then it *still* follows that we're not required to save strangers in rescue cases, which is inconsistent with the Principle of Easy Rescue.

I concede I do not have a decisive argument here that the lives of chickens *can* be aggregated against the lives of humans. There is no canonical criterion for deciding how it applies in a particular case, and the case at hand is not clear cut.[39] This problems run the other way too: I do not expect there to be a decisive argument available that demonstrates the lives of chickens and humans *cannot* be aggregated.

This concludes the analysis of whether abandoning NRAGE makes it easier to hold the Principle of Easy Rescue, given that we accept the Strong Carnism Thesis. It is unlikely that appealing to the direct/indirect distinction changes matters and not obvious that restricting aggregation does either.

4. Can we save the Principle of Easy Rescue?

If we accept the Strong Carnism Thesis and NRAGE, then we must abandon the Principle of Easy Rescue. That we are not required to save lives when we can do so

[39] Voorhoeve, ibid p80 notes that none of (Taurek, 1977; Nagel, 1995; Kamm, 2001) provide precise limits for the personal prerogative that and (Parfit, 2011) claims such matters are 'irredeemably imprecise'.

easily will strike many as counter-intuitive. I propose that, on reflection, we should not find it *very* counter-intuitive.

What might explain our attachment to the Principle of Easy Rescue? The intuitive pull in *Shallow Pond* seems to be based largely on two assumptions: (1) saving lives is overall very good and (2) we have a duty to promote the good when we can do so easily. If we discover saving lives is bad, as it is in *Drowning Dictator*, the intuitive pull disappears. It seemed obvious, prior to accounting for animal impact, that rescuing the person in *Shallow Pond* was very good. Now that we've accounted for the impact meat eaters have on animals, the analysis changes.

What should replace the Principle of Easy Rescue? Given the worry is about animal suffering, it seems it should be:

> *The Anti-Carnist Principle of Easy Rescue*: in rescue cases, we are not required to save the lives of strangers. However, in the unlikely event that we are confident the person is not a meat eater (e.g. they are vegetarian or vegan), we are required to save them.

Few of us are perturbed by the idea that we are not required to save the person in *Dictator Drowning*; if the Strong Carnism Thesis is true, then this case is much more analogous to *Shallow Pond* than we would have thought.

5. Further implications

I note two important implications of the argument.

If the Strong Carnism Thesis is true, saving lives is bad. Thus, for consequentialists and non-consequentialists who hold we are (at least sometimes) required to do the

lesser evil, that implies we're *required* not to save lives in cases of easy rescue.[40] Singer is pushed to say it is wrong to save the drowning child in *Shallow Pond*.

Second, suppose that we, i.e. private individuals, are trying to use our spare money to do as much good as possible. Two obvious options are to give to organisations in the developing world that save lives or reduce poverty. However, the longer people live and the richer they become, the more meat they eat and the more suffering they will cause as a result. Regarding poverty, studies show very consistently that, as people get wealthier, they consume more meat.[41] Hence accounting for human-caused animal impact will reduce the value of both of these by *some amount*—it won't necessarily make such actions negative, not least when we consider that those in the developing world eat much less meat. Crucially, this reduction does not rely on the truth of the Strong Carnism Thesis: just so long as we think humans have *some* animal impact, it applies.

6. Conclusion

Can we believe both that it's wrong to be a meat eater and that we are required to save lives when we can do so easily? I've argued that these two popular, intuitive views are likely to be incompatible, given further assumptions. If it is wrong to eat meat (on animal suffering grounds), then it is certainly plausible meat eaters will cause *enough* suffering that, on different ways of spelling out 'enough', we are not required to save lives. I closed by arguing that accounting for animals' suffering has further practical implications.

---

[40] As an example of a non-consequential who believes we are required to the lesser evil, see (Frowe, 2018).
[41] (Delgado, 2003)

Chapter 2: Saving lives, averting lives, and population size

0. Abstract

Many people seem to accept what I will call the 'Intuitive View', that saving lives is good and—as the Earth is overpopulated—averting new lives is good too. Although it is not widely recognised, the Intuitive View is in tension: population size would also be reduced by not saving lives. I develop Greaves' earlier analysis—which looked at how the Intuitive View could be true—and investigate how probable the Intuitive View is and what the practical implications would be if it were true. I do this first by assuming Totalism and second assuming a Person-Affecting View. I argue that the Intuitive View is distinctly unlikely on Totalism and, further, were it true, the value of each of saving and averting lives would be surprisingly small. This raises problems for Peter Singer' position: he seems to hold the Intuitive View, Totalism, and that life-saving and life-averting charities are among the most cost-effective giving opportunities. On a Person-Affecting view, by contrast, the Intuitive View is far more likely. If it were true the values of saving and averting lives would be in large range—either, but not both, could have around zero value. As such, without further information about how overpopulated the world is, we can say little about the value of either intervention. The general problem the argument makes salient is that we don't know the values of saving or averting lives unless we know how under- or overpopulated the world is. I close by briefly arguing that it's not obvious where the world is in relation to its optimum population size on either a Total or Person-Affecting View.

1. Introduction

Many people believe the Earth is overpopulated: there is a limited amount of food, water, space, and resources available, and the number of humans is now such that it reduces the quality of life for those presently living and those yet to come.[1] Because of this, some people believe it would be better, all things considered, if fewer people were born. Hence interventions that *avert lives*, such as family planning and birth control, are, in general, good.[2]

Hilary Greaves notes that many of the same people who believe the Earth is overpopulated also believe that it is good to save lives.[3] Few, if any, seem to publicly argue saving lives is *bad* (they may privately *think* this), even though the premature deaths of existing people would also reduce population size. As such, many accept what I'll call the 'Intuitive View', the simultaneous belief that both (1) saving lives is, in general, good and (2) averting lives is, in general, good. Peter Singer is a consequentialist who recommends both life-saving and family planning (i.e. life-averting) interventions, implying not only that he accepts the Intuitive View, but that he thinks financing both types of interventions are among the most good individuals can do.[4]

---

[1] For instance, Population Matters is a modern group that campaigns for a 'sustainable population size'. Historical concerns stretch back to (Malthus, 1798) and more recently, (Ehrlich, 1978).
[2] By 'avert lives' I mean prevent lives from ever having started. I am not interested in the question of when exactly life starts and nothing in this essay turns on a particular understanding of this topic.
[3] (Greaves, 2015)
[4] Singer's charitable organisation, (The Life You Can Save, 2019) recommends both charities that save lives, e.g. the Against Malaria Foundation, and those that avert lives through family planning, e.g. Population Services International. For a statement of concern about overpopulation see (Singer, Kissling and Musinguzi, 2018). In writing, Singer advocates saving lives on multiple occasions, e.g. (Singer, 1972, 2009, 2015).

Greaves observes the tension within (what I call) the Intuitive View and raises the question of how it could be true.[5] Greaves investigates the matter assuming Totalism—the population axiology on which the value of a state of affairs is the sum of lifetime well-being of everyone who will ever live—and makes some illuminating conceptual clarifications.[6] While Greaves does not seem to explicitly answer the question she raises, an answer can be inferred from what she writes: the Intuitive View can be true when Earth is at or slightly above its 'optimum population', in which case the value of saving a life just consists in what she calls 'transition factors', the (dis)value associated with the loss of an existing person; I return to and explain these terms later.

In this chapter, I develop Greaves' analysis in two ways. First, I extend it: I address two further questions that are provoked by reflecting on the Intuitive View but not discussed by her. Labelling Greaves' question, 'how could the Intuitive View be true?' as (1), I also consider (2) 'how likely is it that the Intuitive View is true?' (3) 'what are the practical implications of accounting for population size when considering the values of saving and averting lives?' Second, I broaden it: in addition to addressing these questions from a Totalist perspective, I also do so assuming a Person-Affecting View (hereafter 'PAV'). On this PAV, the only people who matter when aggregating individuals' lifetime well-being in order to determine the value of a state of affairs, are those who presently exist and will exist whatever we choose to

[5] All references to Greaves in this paragraph are to (Greaves, 2015).
[6] A population axiology is a ranking of states of affairs in terms of their overall betterness where the number, identities and lifetime well-being individuals of individuals within that state of affairs varies. I am only concerned with axiology (the value of states of affairs) in this chapter.

do.[7] The spirit of such views is captured by 'Narveson's Dictum': "morality is in favour of making people happy but neutral about making happy people."[8]

At this point, it will help to observe that the *overall value* of creating/ending lives is a combination of their *self-regarding* value, the value *solely related* to the person whose life it is, and their *other-regarding* value, the impact that life has on *everyone else*. Clearly, both views (Totalism and PAV) hold there can be other-regarding value in creating/ending lives: births and deaths can impact others. I take it that both views assume there can be self-regarding value in saving lives: it is good if people live longer, assuming they would live happily.[9] For our purposes, the important difference between Totalism and PAV is that Totalists hold that there *is* self-regarding value in creating happy/unhappy lives, whereas PAV holds there is *no* self-regarding value in creating *happy* lives. Advocates of PAV tend to say there is self-regarding (dis)value in creating *unhappy* lives, but I postpone the discussion of this issue until the penultimate section—until then, we are only concerned with creating happy lives in any case. Given this difference in the value of creating lives, the two views will approach the question of optimum population size differently and, as many are sympathetic to PAV, it is valuable to broaden the analysis.

---

[7] This conception of Person-Affecting Views, though imprecise, is sufficient for our purposes; it is borrowed from (Bostrom, 2003) at pp. 311-312. ("Suppose instead that we adopt a 'person-affecting' version of utilitarianism, according to which our obligations are primarily towards currently existing persons and to those persons that will come to exist"). As (Greaves, 2017) section 5 points out, there are a variety of different Person-Affecting Views, all motivated by the intuition there is no value in creating happy lives, that is, lives with positive lifetime well-being (in this context 'happiness' is synonymous with 'well-being'). For summaries of the views and debates in population ethics see the Greaves article just cited and (Arrhenius, unpublished) . I provide a further clarification of different types of Person-Affecting Views in footnote 26 in chapter 3.3.

[8] (Narveson, 1973) p70.

[9] I assume that lifetime well-being is the sum of all the instances of momentary well-being within a person's life. This specification is the Deprivationist View of the badness of death, where death's badness consists in the goods of life it deprives someone of. Alternative views of the badness of death are discussed in chapter 3; their inclusion here would complicate matters substantially without altering the thrust of the analysis.

It has to be the case that, whichever population axiology one uses, the more valuable averting lives is (due to other-regarding concerns about overpopulation), the less valuable saving lives is. I argue that, on Totalism, it is very unlikely the Intuitive View is actually true. The axiological machinery of Totalism, as it holds there is self-regarding value in saving happy lives and self-regarding *dis*value in averting happy lives, simply does not allow much 'room for manoeuvre', such that both saving and averting lives can be, in general, good. This presents an initial problem for Singer as he seems to accept the Intuitive View and Totalism.[10] If the Intuitive View is true, the overall values of both saving and averting lives will be very small (compared to the self-regarding value of saving a life). The practical implication then, if Totalism and the Intuitive View are true, is that neither saving lives nor averting lives seem very promising opportunities to do good. This creates a further challenge for Singer because, as noted, he recommends life-saving and life-averting organisations as being among the most effective giving opportunities.

On PAV, by contrast, it is far easier for the Intuitive View to be true and, if it is, it puts the value of each of saving lives and averting lives somewhere in a large range (I specify this later). Thus, if the PAV advocate knew the Intuitive View was true, but not exactly how overpopulated the Earth is, they would not know much about the relative importance of saving lives or averting lives; they would still know that the more valuable one is, the less valuable the other is.

What emerges from this discussion is that we need to know whether, and to what extent, the Earth is under- or overpopulated to know how valuable it is to save and

---

[10] (Lazari-Radek and Singer, 2014) presents arguments for and against Totalism and conclude on p. 373 that there are deep difficulties in the way of any defensible view on this question. However, in conversation, Singer has stated that, with reservations, he thinks Totalism is correct.

avert lives. In the final part of the chapter, I observe that how one approaches the question of where the world actually is in relation to optimum population will depend on which population axiology one uses: it matters whether we are concerned about the 'near-term', i.e. only this generation, or with the 'long-term', i.e. this and all future generations. However, I then go on to show that determining the reality, on either timescale, of where the Earth is in relation to optimum population is a complex, non-obvious empirical matter.

The chapter is structured as follows. Section 2 introduces Greaves' distinction of the different factors impacting the value of a life and lists some of these factors. Sections 3 to 5 assume, where such definiteness is needed, Totalism. Section 3 defines optimum population (and related terms) and explains the nature of the tri-partite relationship between saving lives, averting lives, and population size. Section 4 gives the answers to questions (1) to (3) stated above. Section 5 anticipates five objections one could make against the argument that, if the Intuitive View is true, the value of saving and averting lives is as small as I claim it is. Section 6 answers questions (1) to (3) on the PAV. Section 7 discusses whether Earth is actually under- or overpopulated. Section 8 concludes.

2. Factors affecting the overall value of life

Greaves draws a helpful distinction between the *absence* and *transition* factors of the overall value of saving a life. Transition factors are defined as those that result from the loss of an existing person. Absence factors are defined as those that result from the *absence* of the latter part of a person's life. The key difference "is that

deprivation factors but not transition factors would still be present if the person in question had never been born at all".[11]

Greaves provides a list of such factors. It is not important to discuss these in great depth, so I will merely paraphrase Greaves' list, labelling absence factors 'A' and transition factors 'T', and noting whether they seem to be positive or negative in value. A4 and A5 are my additions to the list. I note that A1 captures the self-regarding value of saving a life and the others are factors comprising the other-regarding value of doing so. I do not claim this list is exhaustive.

A1 The benefit associated with the saved person living longer.

A2 The positive increase in economic output that results from there being an extra person in society.

A3 The negative environment costs associated with there being an extra person who consumes resources.

A4 The negative impact on non-human animals caused by the consumption of animal products.

A5 The impact on wild animals (i.e. not farmed animals) due to their being an extra person. The thought is that extra humans reduce wild animal habitats, and this will be bad(/good) if the wild animals have, on aggregate, positive(/negative) well-being. It's unclear to me if this factor is positive or negative.

T1 The negative emotional reaction on friends and family as a result of bereavement.

T2 The negative financial loss of having a 'breadwinner' die, where relevant.

[11] (Greaves, 2015) p8

T3 The negative impact on friends and family, once the bereavement effects have subsided, of not being able to socialise with the deceased. As Greaves notes, the counterfactual impact of this is likely to be small and people socialise with others instead.

Armed with these distinctions, we can say that the overall value of saving a life is a combination of (1) the self-regarding value (A1), (2) the other-regarding absence factors (A2-A5), and (3) the value of the other-regarding transition factors (T1-T3). For conciseness, I will refer to (2) and (3) as 'absence value' and 'transition value', omitting reference to the fact they are comprised of other-regarding factors. We can write the value of saving (an existing) life as follows:

Value of saving a life = self-regarding value + absence value + transition value

We can also write the value of averting a new life as follows:

Value of averting a life = - (self-regarding value + absence value)

This is negative, relative to the value of saving a life, as this is the difference in the value of an outcome where we prevent a new person from existing vs they do exist: averting a life means there is one less person's worth of self-regarding value and other-regarding absence impact in the world. Note, importantly, that averting lives lacks the transition value as (by stipulation) that refers to the badness of someone dying once they are alive, which is not applicable to new lives.

3. Relating optimum population to saving and averting lives

In this section, I define optimum population (and related terms), set out the value of saving and averting lives on Totalism at optimum population, and explain what the relationship is between population size, saving lives, and averting lives.

At this point, we should think of what it might mean for the planet to be at optimum population. Greaves proposes the following:

> Suppose further, merely for illustrative purposes, that the world is currently at optimum population, in the following sense: on average, the difference in value between the actual state of affairs and a state of affairs in which a randomly selected actual person does not exist and *others' lives go as they would have if the person in question had not existed* is zero. [emphasis in original][12]

The last part of this is vague: what does it mean for 'other's lives to go as they would have if the person had not existed'? That their lives would have the same lifetime well-being? That they would (implausibly) have led exactly the same lives: had the same jobs, walked down the same streets, and so on?

I suggest alternative, simpler definitions: the world is at *axiological optimum population* when adding new a life has, on average, neutral value. Relatedly then, it is *axiologically overpopulated* when adding a new life is, on average, bad, and *axiologically underpopulated* when this is, on average, good.[13]

The definitions proposed are 'thin' in the sense that they are compatible with any account of population axiology, as opposed to being tied to Totalism in any way: the idea being that, if you, for example, say the Earth is axiologically overpopulated, you

---

[12] Greaves p10.

[13] These claims should account for our expectations of the world's population trajectory, how many will live now and at different points in the future. Note that the Totalist is ultimately concerned with how things go for the 'timeless' population (all those who will ever live) rather than just with how things go for the current population (those alive now). Hence a Totalist making a claim about the world being axiologically overpopulated will consider how adding a life affects the timeless population and not only the current one.

have reached the conclusion, however you think about value, that extra people would be, on average, bad.[14]

Defining overpopulated in axiological terms is somewhat revisionary but essential for our purposes. Sometimes, particularly in environmentalist contexts, discussion of overpopulation is linked to the notion that we (i.e. humans) our exceeding the Earth's 'carrying capacity'; roughly, we are consuming more resources than the Earth can sustainably produce (for a given explication of 'sustainably').[15] Let's call overpopulation in the carrying capacity sense *ecological overpopulation*.

To talk in terms of ecological overpopulation would obscure the issue at hand. For reasons that I give later in section 7, someone could hold the world is *ecologically* overpopulated but *axiologically under*populated. Such a person would not hold the Intuitive View at all.[16]

One might alternatively object to the definitions on the grounds that when we say the Earth is overpopulated, we are not claiming that creating new lives is, on average bad, only that the effects new lives have on others is negative; in our terminology, that the (other-regarding) absence value is negative. To distinguish this term from 'axiological overpopulation', let's say there is 'social overpopulation' when the

---

[14] We could distinguish optimum population for the world from optimum population for some sub-region of it. One might claim the world is overpopulated, adding that a random life is bad; but (say) Japan is underpopulated, then adding a random life in Japan is good. This doesn't remove the challenge of reconciling the parts of the Intuitive View: we can still inquire, for a given sub-region, whether saving and averting lives are both good there.

[15] (Dhondt, 1988) identifies various notions of carrying capacity. These need not detain us here.

[16] To someone who says, 'I think the Earth is ecologically overpopulated but that saving lives and creating lives are good,' we can generate the problematic Intuitive View by replying that 'suppose the Earth is sufficiently ecologically overpopulated that creating lives is bad. How likely is it, in that case, that saving lives is good?'

absence value of adding new lives is negative. We can write this and social underpopulation as follows:

Social overpopulation: absence value < 0

Social underpopulation: absence value > 0

And when social absence impact is zero, we are at the social optimum population:

Social optimum population: absence value = 0

Why not define *axiological* overpopulation in this way? Because it can still be good, on average, to add lives to a world which is *socially* overpopulated. On Totalism, there can be self-regarding value in adding lives. Presumably, the self-regarding value of new lives is, on average, positive—people tend to live happy lives. If the world is *socially* overpopulated—the absence value of extra lives is negative—but the magnitude of the (positive) self-regarding value is greater than the magnitude of the (negative) other-regarding absence value, then adding lives is nevertheless *good overall*. Hence this alternative definition has the same problem as the last one: it doesn't capture the important idea that, when we say the world is axiologically overpopulated, we mean that adding lives is, in general, bad. To show the difference between social overpopulation and axiological overpopulation, we can spell out the mechanics of the latter on Totalism, stated with respect to the value of adding a life[17]:

Axiological Overpopulation: self-regarding value + absence value < 0

Axiological Underpopulation: self-regarding value + absence value > 0

[17] Axiological overpopulation (and cognates) will function identically to social overpopulation (and cognates) on PAV if the view holds there is no value creating happy or unhappy lives.

At Optimum population the overall value of adding a life is zero:

> Axiological Optimum population: self-regarding value + absence value = 0

Having distinguished the 'social' and 'axiological' versions of these terms, I will generally drop the prefix 'axiological' hereafter for stylistic reasons.

If we are at optimum population, what follows regarding the value of saving and averting lives, given Totalism?

At optimum population, adding a new life has zero value. This is because the self-regarding and absence values of adding a life must be *equal* and *opposite* at optimum population. As averting a life is just preventing the addition of a life, averting lives also has zero value.

It also follows that the self-regarding value and absence value of saving lives will be equal and opposite. At least this follows if we assume those values are constant the same through each period of life, which I assume for now and will revisit later. The result is that the value of *saving* a life, at optimum population, will *just* be the transition value, i.e. the effect of the loss of an existing person has on other people.[18] Noting,

> At optimum population: self-regarding value + other-regarding absence impact = 0

Thus,

> Value of saving a life = self-regarding value + absence impact + transition impact

---

[18] I assume the self-regarding value and social absence impact are constant each year. I raise an objection to this in section 5.

As the self-regarding value and other-regarding absence impact cancel out, thus:

Value of saving a life at optimum population = transition impact

There are two points to make here before we move on to the discussion of the Intuitive View in the next section.

First, the transition impact will be very small relative to the self-regarding value of the life (unless we are considering saving very old people). To make this comparison, suppose we are saving some 10-year-old, Xenia, who would live another 80 years. For what follows, it will help to state the self-regarding value of saving a life in 'Well-being Adjusted Life-Years' (WALYs). One WALY equates to the sum total of well-being from one person living for one year at well-being level one. To simplify, I assume all humans we are concerned with live at well-being level one. Thus, the self-regarding value of saving Xenia is 80 WALYs.

In chapter 1.2.2, I supposed that few people consider the transition impact—i.e. mainly the sadness associated with bereavement—to be as much as 10% of the self-regarding value of saving the life. 10% is intuitively too high, as it implies the grief impact is 8 WALYs, i.e. equivalent to 8 years' worth of (net) well-being. In chapter 1.2.2, I also noted that we should assess the transition impact counterfactually: people are generally sad whenever someone dies, hence the disvalue of grief is the difference between the badness now and the badness later. Taking these together, it seems unlikely the transition impact could be more than two WALYs on average. In fact, we can estimate this WALY loss empirically and, on it, two WALYs seems to be

an upper bound; as explaining this estimate is inessential to the argument, I have confined it to a lengthy footnote for the interested reader.[19]

Note that, at optimum population, the value of saving Xenia is just the transition impact, which is two WALYs. The two WALYs figure is 1/40th of the 80 WALY self-regarding value of saving her, which is the component we normally use to judge the badness of a death, and hence it is substantially smaller.[20]

Second, a further concept it will help us to have is 'strong axiological overpopulation' (I use this synonymously with 'strongly axiologically overpopulated'). This occurs when the world is such that the value of saving the life of a *randomly-selected* person is, on average, bad. Thus:

> Strong axiological overpopulation: self-regarding value + other-regarding absence impact + other-regarding transition impact < 0

Or:

---

[19] Let's suppose that self-reported life satisfaction is a reasonable measure of well-being – I defend that it is a reasonable proxy for happiness in chapter 4. (Clark *et al.*, 2018) p81 shows the loss of a spouse causes a roughly 2-point loss of life satisfaction on a 10-point scale over a total of five years – after five years, people seem to have adapted and returned to their pre-loss level of life satisfaction. Hence, the loss is equivalent to someone going from 7/10 to 5/10 for a single year. (Oswald and Powdthavee, 2008) find the loss of a child is about half as bad, in terms of life satisfaction, as the loss of a spouse. Suppose that, on a 0-10 life satisfaction scale, the 'neutral point' equivalent to non-existence is 5. If we assume that an average person has 7/10 life satisfaction, which is about the average score in the developed world, then each year their 'net' life satisfaction score is 2 points (7 – 5). As 1 WALY is equivalent to someone's net well-being for the year, it follows 1 WALY is equivalent to increasing someone's life satisfaction by two points for a year. Hence, the death of a child to one parent has roughly a 0.5 WALYs cost to those parents. We might suppose that the pre-counterfactual loss is about 4 life satisfaction points and thus 2 WALYs – the two parents lose a life satisfaction point overall and some other persons are sad, but less so. When we account for the counterfactuals, this 2 WALYs figure will decrease by some uncertain amount. Claiming that the neutral point on the life satisfaction scale is 5/10 is contentious (I discuss this in chapter 7.3). Arguably, it's much lower. However, if it is lower, that reduces the transition impact. Why is this? Suppose the neutral point is 0, then supposing people have 7/10 life satisfaction, then the *net* annual life satisfaction score, which is equivalent to one WALY, is 7. The resulting 1-point life satisfaction drop parents experience as a result of bereavement is then 0.14 WALYs (1/7) rather 0.5 WALYs, (1/2) and hence the transition value would be around 3.5 times smaller. Hence assuming the neutral point is 5 is the assumption *most generous* to creating a high estimate of the transition impact.

[20] (Greaves, 2015) observes the dominant approach to determining the value of a life is to focus exclusively on the self-regarding value of that life.

> Strong axiological overpopulation: self-regarding value + other-regarding transition impact < other-regarding absence impact

I take this randomly-selected person to be of median age, which means they will have fewer years to live than the average averted person would have had, a consideration which will become relevant later.

4. Examining the Intuitive View

In this section I assess, from the Totalist perspective, questions (1) to (3), namely: can the Intuitive View be true? How likely it is? And what are the practical implications if it is true? I'll show how the Intuitive View is possible, argue it's *unlikely* that it is true, then demonstrate that, if it is true, saving lives and averting lives each do very little good, relative to the self-regarding value of saving a life, and are thus they are unpromising ways to do the most good. In the subsequent section, I consider four objections that an advocate of the Intuitive View could make to show the value of saving and averting lives is higher than I claim here.

One of our earlier definitions, the Intuitive View is true if and only if the actual world is currently above optimum population and below strong overpopulation. It's easy to see why: if the world were not overpopulated it would either be underpopulated or at optimum population, thus averting births would be bad or neutral, respectively. If the world were strongly overpopulated, then it is better *not* to save lives. I take it no one wants to claim (at least openly and in writing), that it is good *not* to save lives, not least because it would imply they have a *pro tanto* reason to stop themselves existing.

> Intuitive view true iff: optimum population < actual world < strong overpopulation.

The problem is, as I will demonstrate, that the difference between a world at optimum population and one where there is strong overpopulation is rather small. Suppose we are considering Tim, who is 30—the world median age—and will live another 40 years—world life expectancy is close to 70. At optimum population, the self-regarding value of saving his life is 40 WALYs and thus the other-regarding absence impact balances it at -40 WALYs. Someone who holds the Intuitive View and believes we are experiencing overall overpopulation must (by definition) believe the absence (dis)value of Tim's life is *already* greater than the self-regarding value of his life. To move from optimum population to *strong* overpopulation only requires the absence value of a life to decrease by the value of the *transition value*. Given the transition value is 2 WALYs, then the world would be strongly overpopulated if the absence value of the rest of Tim's life was -42 WALYs instead.

Whilst it might seem outrageous that the Earth could ever be so overpopulated that saving lives is *bad*, it should be clear that Totalists who believe we are at optimum population should believe, if the absence (dis)value of a life were only *slightly larger*—on these numbers 5% larger—the Earth would be *strongly overpopulated*. I've provided a partial representation of the individual and other-regarding value of saving lives at different population sizes below, in figure 2.1. I've changed the sign of the absence value so it is easier to see where it intersects with optimum population and strong overpopulation.

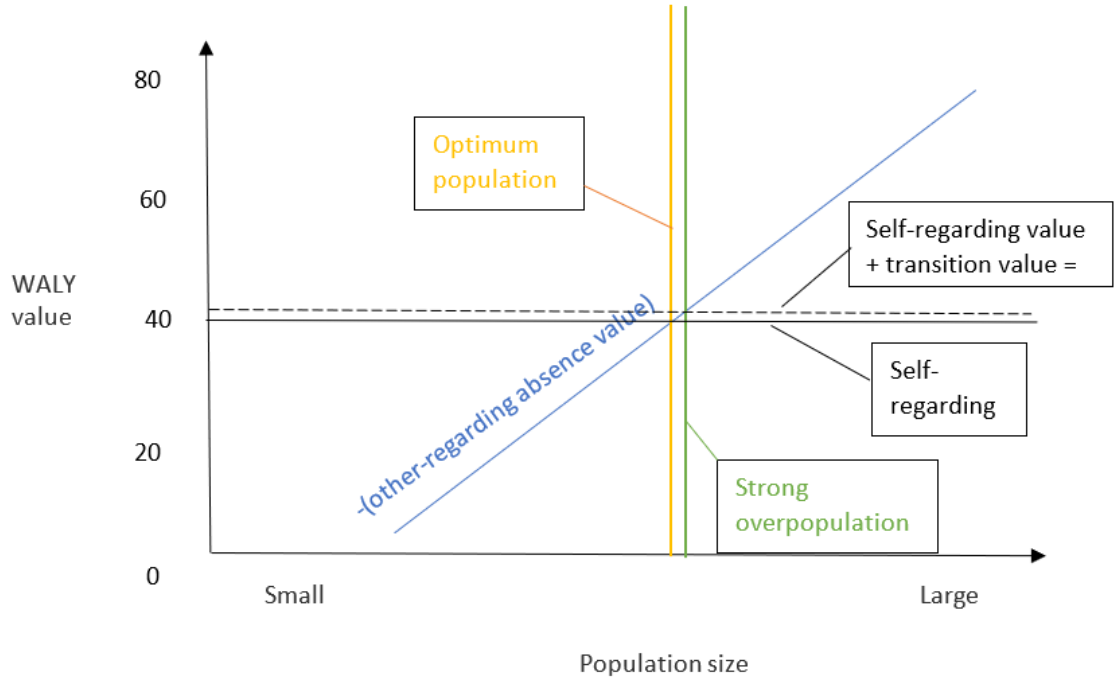

Figure 2.1. Population size vs individual, absence and transition values of saving a median-age life

I return to the question of how under- or overpopulated the Earth actually is later, but it is important to recognise at this juncture how unlikely it is that the Intuitive View is true on Totalism. It would be a striking coincidence if the Earth were in some 'goldilocks' zone: a *little* overpopulated, such that averting lives is good, but not *so* overpopulated that saving lives is bad. It seems much more likely that Earth is either sufficiently overpopulated to make saving lives bad or insufficiently overpopulated to make averting lives bad. Thus, Singer's position—endorsing Totalism and the Intuitive View—is theoretically possible but rather unlikely likely.

Let's consider what the practical implications would be if the Intuitive View were true. The significant result is that the overall value of both saving and averting lives will be far smaller than their self-regarding value and, moreover, those values will be in tension.

Earlier, we specified the values of saving and averting lives as follows:

> Value of saving a life = self-regarding value + absence value + transition value
>
> Value of averting a life = - (self-regarding value + absence value)

For our first-pass analysis, we can assume the life we save and the life we avert are of the same length. If we combine this with our earlier assumption that the self-regarding and absence values are constant each year, then the self-regarding value and other-regarding absence values exactly cancel out we are left with:

> Value of life saved + value of birth averted = transition value

As the transition value is 2 WALYs, what follows is that saving and averting lives are *together* worth 2 WALYs and, further, that the more valuable one is, the less valuable the other becomes. Thus, if both of them are doing *some good*, the amount of good they each do is rather *small* compared to the self-regarding value of saving (randomly selected) lives.

In fact, matters are slightly more complicated. We're talking about saving lives of median age but averting new lives; the latter will be longer than the former. Adjusting for this complicates matters and makes a small difference on Totalism; however, it makes a relatively large difference on PAV and, as it is easier to introduce the analysis now, I will do so.

We can say (using the same numbers as above):

> Value of saving a median life = self-regarding value + absence value + transition value
>
> = 40 + X + 2
>
> = 42 + X

Where X represents the absence value, which changes depending on how under- or overpopulated the Earth is. If the Intuitive View is true, then X must be between -40 WALYs (optimum population) and -42 WALYs (strong overpopulation). The value of averting a new life, relative to the value of saving the median life is:

> Value of averting a life, relative to value of saving a median life = - (self-regarding value of saving median life + absence value of saving median life) * (life expectancy of new lives/ life expectancy of median age person)

This accounts for the fact the new life would live proportionally more years than the median life. Plugging in the numbers in this case, assuming a new person lives to 70:

> Value of averting a life, relative to value of saving a random life = (70/40) * -(40 + X)
>
> = - (70 + 1.75X)

We can combine the two equations to show what the value of saving one random life and averting one life would be:

> Value of saving a random life and value of averting a life = 42+ X – (70 + 1.75X)
>
> = -28 + 0.75X

At optimum population (i.e. X is 40), the value of saving a random life is 0 WALYs, the value of averting a life is 2 WALYs. At strong overpopulation (i.e. X is 42), the value of saving a life is 0 and the value of averting a life is 3.5 WALYs. Hence, the value of saving one life and averting one life is *not* just equal to the 2 WALYs transition value of saving a life, but it isn't far off (given the Intuitive View). Figure 2.2 plots the values of saving a median life and averting a life for different absence

values and indicates the clear tension that exists between the value of saving and averting lives.

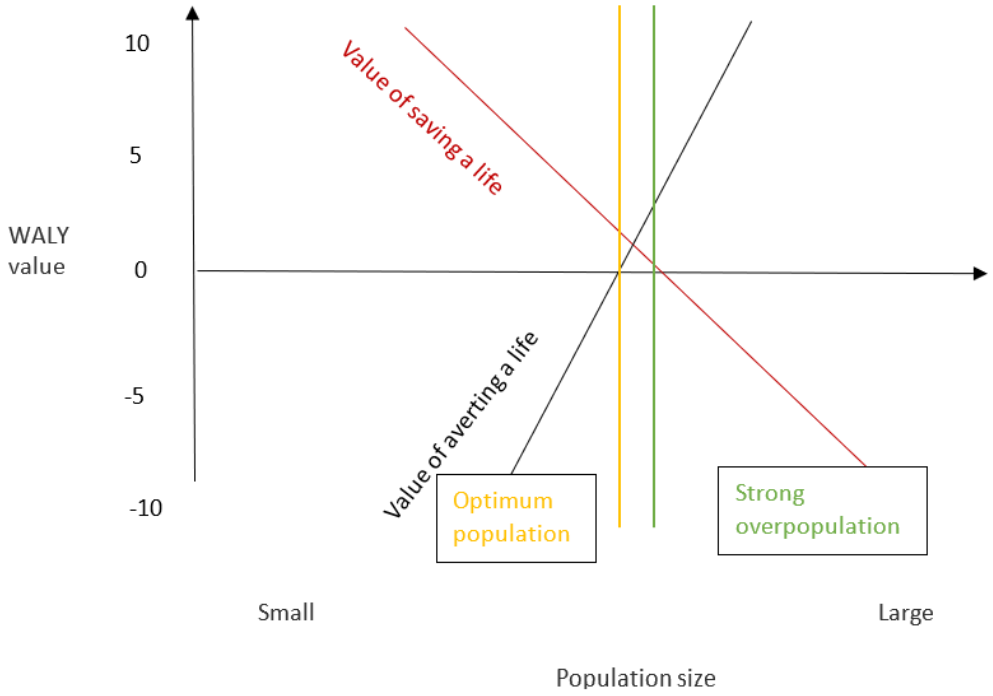

Figure 2.2 Population size vs value of saving or averting lives on Totalism

Now that we have shown the value of saving and averting lives is small, relative to the self-regarding value of saving a life, we can indicate a further problem for Singer's position, given that he recommends life-saving and life-averting charities as being among the most cost-effective.[21]

As far as I can tell, Singer has not stated that his recommendations to save lives take account of his (elsewhere mentioned) concerns about overpopulation. Let's assume

[21] See footnote 4.

Singer's suggestions have not accounted for absence (dis)value.[22] Incorporating these concerns make a big difference. To see this, we can suppose a child saved by the Against Malaria Foundation—a charity Singer endorses and which provides anti-malarial bednets—would live 60 more years, and thus the overall value of saving that life, ignoring the absence value, would be 62 WALYs—60 for self-regarding value and 2 for transition value. Let's assume the Earth is at optimum population and we now incorporate the absence value of that life, which will be -60 WALYs. When we do this, the overall value of saving the life drops to just 2 WALYs, the transition value, which is over 30 times smaller than the self-regarding value of doing so and what I assume Singer's recommendations were based on. Thus, unless we already thought that saving lives in this way was 30 times *more* cost-effective than the next-best altruistic options—e.g. alleviating poverty, treating mental illness, reducing factory farming, etc.—saving lives would *not* to be most good we could do.

In fact, the value of saving lives will be less than this if the Intuitive View is true, as the Intuitive View doesn't entail that the Earth is *at* optimum population, but that it is somewhere *above* optimum population and *below* strong overpopulation. This means the value of saving that life is *between* 0 and 2 WALYs – even lower than the figure used above of 2 WALYs.[23]

Therefore, it does not seem credible, if Totalism and the Intuitive View are true, to hold saving lives or averting lives will be the most good one could do. At least, it is not credible for someone to hold this unless (a) they have acknowledged how much

---

[22] The other less probable explanation would be that Singer has factored in concerns about population size but not mentioned them anywhere.
[23] I won't give an example, but I note that life-averting interventions, whatever we thought their value was before, will be between 0 and 3.5 WALYs each and thus are still relatively small.

smaller the value of saving and averting lives would be once the absence value is accounted for and (b) they show that, even including this reduction, the value of one of saving or averting lives is still better than the alternatives.

It might seem that I am being unreasonably critical of Singer's position. It's worth adding the Intuitive View poses something of a trilemma for Totalists. Option (1), the one I've discussed, is that the Intuitive View is true. As noted, this seems unlikely to be the case and, if it is, neither saving nor averting lives would be particularly valuable. Option (2) is that the Intuitive View is false because saving lives is, in general, bad, i.e. the Earth is not merely overpopulated, but strongly overpopulated; this would make averting lives more valuable but is, presumably, an unwelcome result. Option (3) is that the Intuitive View is false because averting lives is, in general, bad; this would mean saving lives is relatively more valuable but also that the Earth is *under*populated. If the Earth is underpopulated then, at least *prima facie*, it would be better, *inter alia*, if government stopped funding family planning (i.e. contraceptives) and encouraged people to have larger families instead. Therefore, all the options available to the Totalist are at least somewhat counter-intuitive. To state the obvious, the question of how under- or overpopulated the Earth is, once one's axiological machinery is specified, is an empirical question, and it would be *ad hoc* to decide how under/overpopulated the world is based on what seems to have the least counter-intuitive implications.

5. Five ways to increase the value of saving/averting lives, given the Intuitive View

I imagine those attracted to the Intuitive View will think this has all been too quick. I'll now consider five ways someone could respond if they found the Intuitive View plausible but thought that the value of saving and averting lives were both positive

and could be larger than I claimed. For concreteness, I'll suppose the aim is to hold the Intuitive View and restore the idea that the overall value of saving a life is roughly the same as the self-regarding value of doing so. I argue that none of the responses succeeds in achieving this aim.

5.1 Claim the transition costs are higher

The first move is to claim the transition value—the badness to the living of losing someone once they exist—is much higher than I estimated. As the value of saving a life and averting a birth is together is a function of the transition value of saving a life, if this value is much higher, that gives those tempted by the Intuitive View more 'room for manoeuvre'.

For this objection to succeed, the transition value would need not just to be a bit higher than I suggested, but as large the self-regarding value of saving a life. This doesn't seem plausible: generally, when we think about the badness of someone dying, we assume the loss to them is going to be far greater than the sadness we'll feel at losing them.[24] Further, even if the transition value were higher, at say 10 WALYs, then the argument of the previous section—that the Intuitive View is unlikely on Totalism—stands.

5.2 Claim Earth is only socially overpopulated

The second move is to drop the claim that the Earth is *axiologically* overpopulated to the more modest one that it is *socially* overpopulated. As such, there would still

[24] The most promising way to make this claim would be to hold the preference-satisfactionist of theory of well-being. On this, well-being consists in one's preferences being satisfied, in the sense that what one prefers to be the case, is the case. One could then argue that losing one's child, partner or friend is something one very much prefers was not the case and thus represents a large loss in well-being.

be absence disvalue associated with lives, but this is of a smaller magnitude than the self-regarding value. Suppose the absence value that Tim, our median man, causes by living 40 more years is equivalent to -20 WALYs instead of -40 WALYs. This allows advocates of the Intuitive View to retain their belief that there is *in some sense* too many people whilst increasing the value of saving lives, in this case, from 2 WALY to 22 WALYs.

The problem with this objection is that, if the Earth is now axiologically *under*populated, the Intuitive View is false.

5.3 Claim averted lives have lower well-being than saved lives

Earlier, I assumed that saved and averted lives would have the same well-being. However, one might object that this is mistaken on the grounds that the averted lives would generally have lower average well-being than average lives (in a given country). The rationale is that parents who would, on the provision of family planning services, choose not to have children make such a choice, at least in part, because they think that their prospective child would be born in unfortunate circumstances and they wish to avoid this.[25] As such, someone could hold the Intuitive View and clarify they think saving and averting *average* lives is good. They could then claim that (a) life-averting interventions are quite valuable in reality as the averted lives would be of *unusually* unhappy people and (b) life-saving interventions, in contrast, will reach people of average well-being.

Two comments. First, while one way to reduce the number of new births is to provide contraceptives to prevent *unwanted* children, the other obvious option is to

[25] If averting lives occurred through coercion, e.g. forced sterilisation, it would no longer be plausible those averted children would have lower well-being.

reduce the number of children people *want* through, e.g. educating women and girls and changing their preferences regarding an ideal family size. Suppose some parents had wanted five children but then decided they would prefer to have three; it's not clear whether the 'extra' three children would have had particularly worse lives than the first two. It's just that the parents, considering their own interests, changed their minds. Hence, it's not clear, from the armchair, whether the lives averted in practice would have lower average well-being than average new lives.

Second, even if averted lives would, in practice, have lower well-being than average lives, presumably this difference is slight. Suppose averted lives have 10% lower average well-being. In this case, the value of averting a 70-year life, at optimum population, would be 7 WALYs rather than 0 WALYs (the figure of averted lives had average well-being); thus, this makes the Intuitive View more likely, but still unlikely.[26] Moreover, the averted lives would need to have net zero (or net negative) well-being to make it the case that the overall value of saving/averting lives is around the self-regarding value of saving a life; this seems implausible.

5.4 Claim saving lives reduces population size

The third option is to point out that saving lives can often cause people to have fewer children. We call this the 'save lives to avert lives', or 'SLAL', approach. This would seem to be a two-for-one bargain for advocates of the Intuitive View: you get to save lives while averting lives, both of which the view holds are good. The explanation for how SLAL works is that parents, at least in the developing world, have more children than their ideal family size in the expectation some will not survive to maturity. Therefore, if lifesaving interventions are provided, and parents see these working,

[26] The self-regarding value is 0.9*70 = 63 WALYs.

either on their own children or in the community, they will opt to have fewer children instead.[27] We can call this phenomenon the 'reduction effect' and use the term 'reduction rate'—hereafter 'RR'—to refer to the *number of births averted per life saved*.[28]

I will explain how SLAL works, then give two reasons why this doesn't increase the value of saving lives substantially above their transition value. To do this, it will help to represent how SLAL works mathematically:

> Value of the reduction effect: value of saving a life + (RR)(value of averting a birth)

Expanding the component terms, we can say the value of the reduction effect is *positive* when:

> Self-regarding value + absence value + transition value – (RR)(self-regarding value + absence value) > 0

For simplicity, I'll assume the saved and averted life are the same length.[29] We can combine these values and rewrite it as:

> (1 – RR)(self-regarding value + absence value) + transition value > 0

Let's plug in some numbers to illustrate this. Suppose RR is 2, i.e. one life saved averts two births, then the self-regarding value of saving a life is 40 WALYs, the

---

[27] (Doepke, 2005)

[28] It's worth noting that philosophers have debated the moral value of replacement compared to life extension, e.g. (Arrhenius, 2008). Outside philosophy, Melinda Gates of the Gates Foundation has offered this as a rationale of why saving lives is good but doesn't cause overpopulation (Melinda Gates, 2014).

[29] As noted earlier, this isn't very distortionary on Totalism, even if the saved life is of median age. It's even less distortionary in this case as we're mainly considering saving the lives of young children, whose life expectancy will be only a couple of years more than that of new lives.

transition value is 2 WALYs, and the absence value is -42 WALYs, i.e. we are on the cusp of strong overpopulation. This means:

$$= (1 - 2)(40 - 42) + 2$$

$$= (-1)(-1) + 2$$

$$= 2 + 2$$

$$= 4$$

Saving one life and averting two lives is good in this case and has an overall value of 4 WALYs.

The first reason invoking SLAL doesn't make much difference, as the above equation shows. This reason is that if RR is one, then the value of saving each life is just the transition value. As a matter of fact, RR might be around one: GiveWell estimate that for each life the donors to their charities are able to save, about one less person is born.[30]

The second reason is that if the Intuitive View is true, the *maximum* value for each life averted will be the transition value: at strong overpopulation, the value of averting a life is equal to the transition value of saving a life. Thus, even in the implausible scenario that saving one life averts four lives, the value of doing this would only be 10 WALYs, five times the transition value of 2 WALYs. Of course, SLAL would be more valuable if the Earth were strongly overpopulated, but then the Intuitive View would be false: if the Earth is strongly overpopulated, then saving lives *outside* SLAL scenarios would be bad.

---

[30] (GiveWell, 2014)

5.5 Claim value of a life varies with age

Fourth, we could reject my simplifying assumption that the annual absence value of a life is constant for each year of life. Thinking economically, people are a cost in the first part of their life because society invests in their education, healthcare, etc. People are economically productive thereafter and stop being so when they retire. Hence, one might claim if the world were at optimum population, even though creating a new life would have (in expectation) no value, saving the life of an existing 20-year-old's life is more valuable and could be far more valuable than just the transition value. The value of a life at different ages might vary roughly along the lines represented below in figure 2.3.

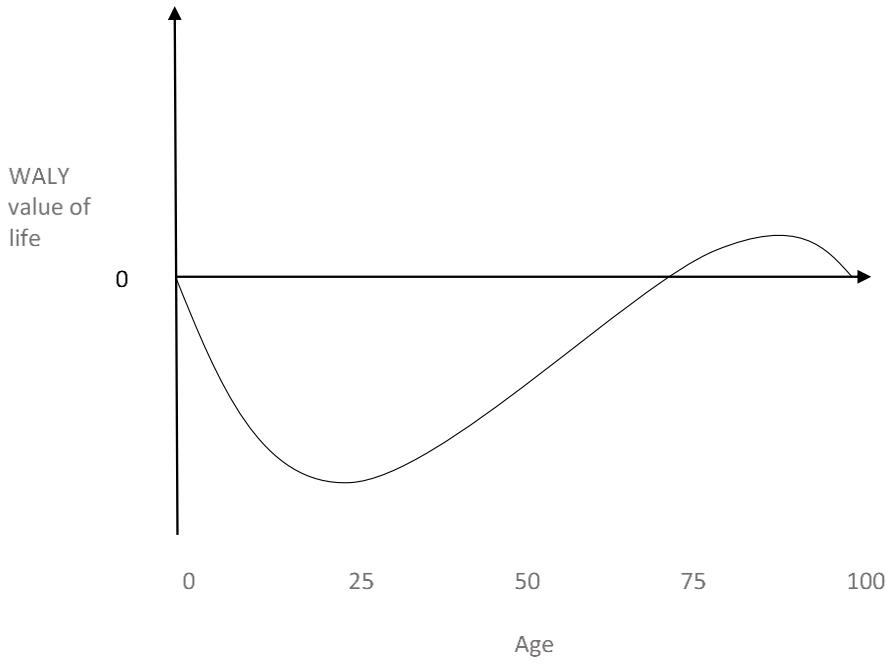

Figure 2.3. Potential variable value of a life over time

This implies it's good to save someone's life up until the point they retire (I've marked that to be around 70) as the counterfactual difference between those negative values (e.g. death at 25) and zero (i.e. death at 100) is positive. It also implies it's good if people die just after retirement and it's good to let young people

die if they aren't going to live long enough to be productive (i.e. it's bad to save a 10-year old who would die at 20).

While it is plausible that the value of lives varies with age, the problem is knowing what to put on the y-axis, which I've intentionally left blank. Is someone's working life from age 25 until their death worth 1 WALY to others? 10? 100? As I discuss in section 7, the sign and magnitude of the other-regarding impact of lives are unclear. Hence, it's not obvious how strong this objection is. It would be *ad hoc* to, without further investigation, put a number on this.

To sum up, none of these five objections succeeds in restoring the idea that the overall value of saving or averting lives can be large—specifically, as large as the self-regarding value of saving a life—if one accepts Totalism and the Intuitive View.

6. PAV and the Intuitive View

In this section, I tackle the same questions as I did in section 4, this time from the perspective of the PAV. It would be tedious and unnecessary to repeat the earlier analysis at length because, as noted in section 2, the only difference between PAV and Totalism is that the latter holds there can be value in creating new lives, and the former holds there cannot be value in creating happy lives. This section continues to assume that all the lives we are considering creating would be happy.

The Intuitive View will be true on PAV under the same description—when the Earth is above optimum population but below strong overpopulation. The difference is that, on PAV, as there is no self-regarding value in adding lives, the world is overpopulated if the absence value of adding a life is negative. Thus, when we consider which factors they incorporate, denoting the population axiology with a

subscript, axiological overpopulation for the two views are to be understood as follows:

Axiological Overpopulation$_{\text{Totalism}}$: self-regarding value + absence value < 0

Axiological Overpopulation$_{\text{PAV}}$: absence value < 0

As such, on PAV, axiological overpopulation functions identically to, but is conceptually distinct from, social overpopulation.

Regarding the Intuitive View, what follows is that it is much easier on PAV for the Intuitive View to be true.

Let's return to Tim, our imaginary median-age person, who is 30 and has 40 years to live. The self-regarding value of saving him would be 40 WALYs in self-regarding value and 2 WALYs in transition value. At optimum population, the absence value of the remaining 40 years of Dan's life must be 0 WALYs. Thus, saving him is worth 42 WALYs. To get to the point of strong overpopulation, the absence value of saving Dan would need to decrease from 0 WALYs to -42 WALYs. Recall, by contrast, that in order to get from optimum population to strong overpopulation on Totalism, Tim's absence value had to decrease by 2 WALYs, from -40 to -42. Hence, as a proportion of his self-regarding value, Tim's absence value needs to increase by 5% on Totalism, but 105% on PAV, for the world to move from optimum population to strong overpopulation. Thus, to state what is transparent, it is far more likely for the Intuitive View to be true on PAV. This is not to say the Intuitive View is necessarily true: the world could still be underpopulated or strongly overpopulated.

What are the practical implications of this on PAV if the Intuitive View is true? What remains the same is that the more valuable saving lives is, the less valuable averting lives is (and vice versa); what changes is that the values of saving and averting lives

are in quite a large range. Therefore, if a PAV-advocate knew that the Intuitive View is true, but not how overpopulated the world is, they would not know enough about the value of saving or averting lives to conclude which of those two options was better.

Slightly adjusting the analysis from section 4, let's state the value of saving a median life and averting a life.

Value of saving a median life$_{PAv}$ = 40 + X + 2

= 42 + X

Where X represents the absence value. If the Intuitive View is true, X is between 0 WALYs (optimum population) and -42 WALYs (strong overpopulation). The value of averting a new life, relative to the value of saving the median life is:

Value of averting a life relative to the value of saving a random life$_{PAV}$ = - (absence value of saving median life) * (life expectancy of new lives/expected remaining life expectancy of median person)

Note this is different from the Totalism's equivalent as, on PAV, there is a self-regarding value in averting lives.

Plugging in the numbers in this case, assuming a new person lives to 70:

Value of averting a life relative to the value of saving a random life$_{PAV}$ = -(X) * (70/40)

= - 1.75X

We can combine the two equations to show what the value of saving one random life and averting one life would be:

Value saving a random life and value of averting a life$_{PAV}$ = 42+ X + (-1.75X)

= 42 - 0.75X

If X is 0 (i.e. at optimum population), the value of saving a random life is 42 WALYs and the value of averting a life is 0. If X is -42 (i.e. at strong overpopulation) and the value of saving a life is 0 and the value of averting a life is 73.5 WALYs. Thus, the value of averting a life at strong overpopulation is much larger than that of saving a life at optimum population; this is a result of the fact the averted life would live much longer than the saved life would. This is shown in figure 4; note the trade-off between the value of saving and averting lives remains.

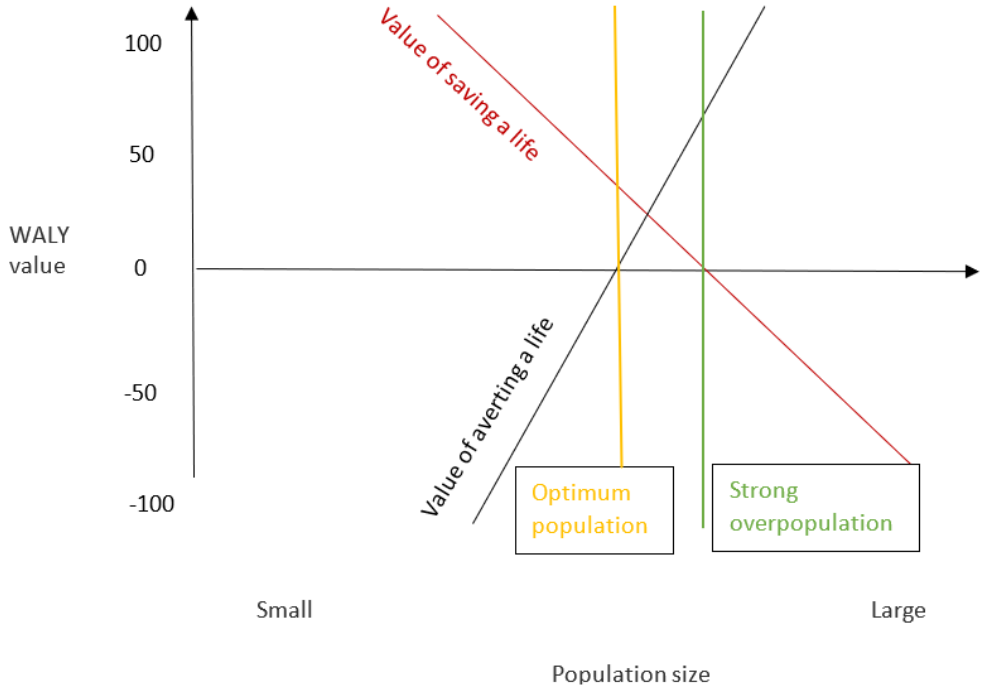

Figure 2.4. Population size vs value of saving and averting lives on PAV; note the identical gradients but different scales to those in figure 2.2.

Suppose someone advocates PAV, endorses the Intuitive View, and recommends that either saving lives or averting lives is the most good one can do. Suppose, further, that such a person has not yet factored in the absence value of lives. Factoring this in would necessarily reduce the value of saving lives and increase the

value of averting lives. By what proportion would this reduce the overall value of saving lives? 10%? 50%? 90%? It is important to realise that if one endorses the Intuitive View but cannot say with any precision how overpopulated the Earth is, one would not know which of saving or averting lives is better: on the Intuitive View, the value of one of saving or averting lives (but not both) could be zero. The more general point is that, regardless of whether one accepts the Intuitive View, unless we know the absence value of lives, we do know how valuable saving or averting lives is.

A temptation here would be for such a person, in the absence of any further evidence, to arbitrarily pick some reduction. Perhaps they would assume we are halfway between optimum population and strong overpopulation. To state the obvious: to do this, without evidence, would be *ad hoc*. The prudent move would be to try and form some empirical-informed estimate, which is what we discuss and fail to do in the next section.[31]

7. Optimum population and population ethics

The preceding sections have demonstrated that much turns on whether, and to what extent, the world is in fact under- or overpopulated. In this section, I argue (perhaps disappointingly) that determining this is complicated—too complicated to tackle here—and it is not at all clear what the reality is. There are two parts to this claim.

---

[31] Regarding the five objections in section five, the only one that functions differently between Totalism and PAV, and is thus worth mentioning, is number three, the 'SLAL' approach. On PAV, SLAL is rather more appealing as it doesn't count it as bad that the averted persons don't get to live happy lives. The formula for the value of the reduction effect is as follows:

Value of the reduction effect: value of saving a life + (RR)(value of averting a birth)
= Self-regarding value + absence value + transition value – (RR)(absence value)
= Self-regarding value + transition value + (1-RR)(absence value)

If RR is 1, then the value of saving a life and averting a life is equal to the self-regarding value and transition value of saving one life, i.e. equal to the overall value of saving a life at optimum population. Again, this assumes that the saved and averted lives are of the same length.

First, I show how the relevant time span to consider, when we are thinking about population size, turns on which theory of population ethics is correct. Second, I sketch some ways of assessing optimum population and show it is non-obvious what the reality is from either a near- or long-term perspective.

How optimum population is assessed will depend on which population axiology is used. Totalists hold all possible lives matter and are therefore concerned in principle about the 'long-term', i.e. effects on this and all future generations. By contrast, on PAV, at least *prima facie*, we are concerned only about the people who currently exist and will exist, whatever we do. On the plausible assumption that our actions now will eventually change the identities of everyone who ever lives, and that the number of people who will exist 'whatever we do' is relatively small compared to those who currently exist, PAV is, in effect, primarily focused on the 'near-term', i.e. the effects on the current generation.[32]

I say 'prima facie' regarding PAV because, as noted, many advocates of PAV endorse the 'procreative asymmetry', according to which the addition of unhappy lives makes the world worse (the 'first conjunct'), while the addition of happy lives does not make the world better (the 'second conjunct').[33] An 'asymmetric' PAV is one which accepts both conjuncts of the procreative asymmetry; a 'symmetric' PAV rejects the first conjunct, i.e. adding lives (happy or unhappy) *does not* make the

---

[32] (Parfit, 1984) at ch. 16 (convincingly) argues that the people who will get born will be altered if the parents meet and have children at even slightly different times – a different sperm and egg combination will meet and so a genetically different person will be born. Clearly, you would be someone else if you had different genetics. Even small changes will 'ripple' through society and eventually change all future identities. To aid the disbelievers he offers (on p. 361) the following: parenthetic remark "It may help to think about this question: how many of us could truly claim, 'Even if railways and motor cars had never been invented, I would still have been born'?"

[33] For discussion see e.g. (McMahan, 2009).

world go better or worse. Totalism is an example of a view which rejects the second conjunct.

While symmetric PAVs will focus on the near-term, this is not *necessarily* true for asymmetric PAVs. What complicates matters is there are two ways of cashing out the procreative asymmetry when it comes to adding a 'mixed bag' of lives, i.e. where some will be happy and others unhappy. On the 'strict' version of the asymmetry, we ignore the happy lives in this mixed bag and count the unhappy lives. As Beckstead observes, the strict view is remarkably, indeed implausibly, strict: it would count it as bad to add a billion happy lives and one unhappy life.[34] Given how many unhappy lives there could be in the future—particularly if we expand our thinking to include non-human animals as well as animal—as the strict view will hold these unhappy lives matter, it will be focused, as Totalism is, on what happens over the long-term, albeit just on reducing the quantity and increasing the quality of these unhappy future lives.

On a 'neutralising' version of the asymmetry, when we add a mixed bag of lives, the happy lives can 'cancel out' the unhappy lives, in the sense that the happy lives reduce the extent to which the unhappy lives make the world go worse.[35] We can suppose, for definiteness, this works as follows: if the sum of well-being of the additional lives is neutral or positive, adding them is neutral in value; if the sum of well-being of the additional lives is negative, adding them is negative in value (and further, the negative value is equal to the sum of (negative) well-being). Thus, the addition of a billion happy lives and one unhappy life would be counted as neutral

---

[34] (Nicholas Beckstead, 2013) at pp.82-6.
[35] (Frick, 2017) endorses such a view, calling it the 'principle of holistic neutrality'.

in value (unless the unhappy life was a billion or more times worse than the average happy life was good).

What does this view imply? Suppose you expect the future to be sufficiently good that, *whatever you do*, you cannot reduce the sum of well-being of the *additional* lives to such an extent that it becomes negative. In this case, you are axiologically impotent regarding additional lives—you cannot make things go better or worse as it pertains to them.[36] The neutralising asymmetric PAV would then function, in practice, identically to a symmetric PAV—the focus, insofar as one aims to do good, is on lives in the near-term. If, on the other hand, the value of the additional lives is negative and you can do something to reduce the sum of negative well-being, then you will not *only* be concerned with the near-term. I take no stand on what the future looks like or whether we can affect it: my point is only to observe that, on the neutralising version of the procreative asymmetry, the time-frame that is relevant is sensitive to such further considerations.

We now come to the second part of the section—whether and to what extent the Earth is under or overpopulated, given one is taking a near- or long-term perspective. When faced with a tricky problem, a standard way to make progress is to make some simplifying assumptions and see where that gets us. What I do here is to introduce a couple of relatively simple models for assessing this question and show it's unclear on either of these where we are in relation to optimum population (from either the near- or long-term perspective). I then point out the second model

[36] A perverse implication of this is that it would count as neutral to reduce the sum of well-being in the mixed bag by e.g. swapping some good lives for bad lives, so long as the overall bag was not negative in well-being.

is still far too simple—it omits various important considerations and their inclusion does not make the answer to the question at hand any clearer.

If we want to get our teeth stuck into the issue of optimum population, a reasonable place to start is by assuming that the Earth is overpopulated if and when it exceeds its 'carrying capacity', the maximum population size (of a given species) that an environment can support indefinitely (or, at least, for the foreseeable future). If the limit is exceeded, the death rate must *eventually* increase by, for example, famine or war to bring the numbers back down.[37] Given the suffering involved, we might expect a wide consensus on different ethical views that it is best that the Earth's population avoids going over this limit (assuming it is not already above it).

As Greaves points out, citing a survey by Cohen, the issue with appealing to the carrying capacity arguments is that there is a distinct lack of consensus on what the carrying capacity is: of Cohen's survey of 65 estimates, half lie in the 5-14 billion people range, and a third are above 20 billion.[38] Of course, if one expects the Earth's population to grow *ad infinitum*, one would worry we will eventually hit this 'ceiling', whatever it is. However, it seems unlikely that the Earth's population will increase very dramatically. As countries develop, people have fewer children. In all European countries, fertility is now below the long-term replacement rate—the number of children per woman required to maintain population levels—of around 2.1 children per woman and in many European countries, it has been below this for decades.[39] While 21% of the world's population was in societies with below

---

[37] A further issue is whether exceeding carrying capacity does or does not lead a permanent reduction in the maximum number of lives supportable by the environment. For simplicity, I set this to one side.

[38] (Greaves, no date) at 3.1 cites (Cohen, 1995) Ch 11. and Appendix 3.

[39] (United Nation Department of Economic and Social Affairs, 2017) at p6.

replacement fertility rates in 1975-80, this figure is estimated to be 69% by 2045-50.[40] Extrapolating this suggests that the population will eventually stabilise and then decline, rather than increase indefinitely. From UN estimates, it seems that the world's population will probably come close to its peak around 2100 at just over 11 billion people.[41]

Hence, to make the case we are or will crash into the carrying capacity ceiling, one would need to get 'into the weeds' of the estimates and show a lower estimate that is more plausible. As Greaves also points out, making sensible arguments here is tricky. This is because, as she notes, carrying capacity is not fixed, at least for humans, but depends instead on the ability of technology to support a larger population by, e.g. developing a more efficient farming method to avoid famine. While one might think that technological development is independent of population size, on a 'Boserupian' view of innovation—named after the economist Ester Boserup—a larger population size causally leads to improved technology.[42] On this view, necessity is the mother of invention, and it is exactly the pressures that larger populations put on society which cause the inventions that allow carrying capacity to increase. Accounting for the Boserupian view does not, however, mean we should ignore the environment degradation larger populations causes, only that it is not the whole of the analysis when considering carrying capacity, and thus that carrying capacity is likely to be larger than we would think if we had not accounted for it.

Leaving aside these complexities, even if we could nail down the Earth's carrying capacity, this is, at most, only one part of the puzzle. Near-termists won't necessarily

---

[40] Ibid, p7.
[41] Ibid, figure 2 at p2.
[42] E.g. (Boserup, 1981).

mind that we are or will be hitting the ceiling—the problems may befall later generations who are deemed to be morally unimportant. Equally, long-termists could have reasons for exceeding carrying capacity—perhaps a larger population means a larger economy which, in turn, makes it easier to colonise the galaxy, something that is plausibly of enormous value. More generally, both near- and long-termists will care not just about the quantity of supportable lives—which is what carrying capacity refers to—but the quality of those lives.

This brings us to our second type of model, an economic one that accounts for the relationship between population size and well-being levels. To get this going, let's grant or note the following. First, well-being is a function of consumption.[43] Second, individuals produce and consume resources. Third, the average resources produced per head follows something like an inverted-U shape. To explain the third assumption, the idea is that at low levels of population, economics of scale are not possible—a certain population size is needed to support valuable specialised professionals, e.g. doctors, engineers, metaphysicians, etc. However, at very high levels of population, diseconomies of scale emerge: certain resources, such as land, run out, pollution and congestion become problematic, and so on. What follows is that each additional person produces some resources, consumes some resources, and (presumably) experiences positive well-being. Maximum well-being among the 'momentary' population (those existing at a given moment) will be reached when these factors balance: the additional person converts the resources they consume into the same sum of well-being that this person removes for others by reducing their average consumption. (Recalling the earlier discussion, this is merely a

[43] It need not be a function only of consumption, but assuming this would further simplify things.

different way of saying the self-regarding value of the life and the absence value are both equal and opposite.)

While it is not necessary to get into the details here, it's possible to construct a mathematical model using these assumptions on which, if one knows (a) the 'production function', the relationship between population size and production and (b) the 'consumption function', the relationship between consumption and well-being, these numbers can be crunched to determine which population would be optimal to maximise the total well-being of the momentary population.[44] The major problem here is that we need to supply the production and consumption functions as inputs to the model. Determining what these are is a complicated empirical question, one that, as far as I know, no one has yet tried to tackle and I am unable to undertake here. I note that tackling it involves assessing some of the problems raised in the previous model about how population size affects the environment and technological development.

Even this model does not capture all the considerations that seem relevant to long-termists. It captures the momentary optimum population, when what is wanted is a model of optimum population across all generations. Long-termists will likely want to ensure that (a) humanity does not become extinct and (b) it colonises space.[45] It is not clear how one would connect the economic model of population size to (a) or (b) or, if one did, whether a larger or smaller population would be better.[46]

---

[44] See (Greaves, no date) footnote 14 for such number crunching. The population that maximised average well-being could also be calculated, but that is not our concern here.
[45] For an expression of such a view, see (Bostrom, 2003), (Greaves and Ord, 2017), (Beckstead, 2013).
[46] Let's illustrate this regarding space colonisation. A smaller population would seem on the one hand to be better: it would cause less environmental strain and allow us to survive within Earth's ecological limits for longer, and provide us with more time to perfect space-faring technology; on the other hand, a larger population would mean a larger economy, and thus more resources would be available for space-faring projects.

Where does this leave us? In short, it's unclear whether the Earth is under or over-populated on either a near or long-term perspective.

At least, this is the case if we only count the impact that humans have only on *other humans*. What's been left out of the analysis so far is the impacts humans have on non-human animals, notably those that are caused to exist in factory farms by the demand for meat and may well have unhappy lives—this was the subject of the previous chapter. [47] Given how difficult it is to estimate the sign and magnitude of the effect of humans on humans, a temptation is to ignore this and consider the easier-to-estimate impacts on animals. While this is tempting, doing so would leave us open to accusations of quantification bias, as illustrated by the old joke about the person looking for her keys under the lamppost, not because that's where she lost them, but because that's where the light is (and thus the only place she can look). It's unclear how reasonable it is to ignore factors which we accept might be very large but we know neither the sign nor magnitude of, and just account instead for factors we believe are potentially rather small. It seems we should invest our resources in trying to understand the effects that humans have on other humans—increasing how far the lamppost's light shines, as it were.

While making progress on this question is complicated, it does not seem vastly more complicated than other complex modelling exercises, for instance in macro-economics or climate science. This seems (yet) another case where further detailed research is required. As Greaves notes at the end of her analysis of optimum population, the 'more research, please' cliché is at least slightly more interesting in

---

[47] There's also the question of the well-being of wild animals, which is too complicated to discuss here. For a discussion of this, see e.g. (Tomasik, 2015) who argues there is large net negative well-being in nature.

this case than it normally is: participants in debates on population size either take it to be obvious that the Earth is overpopulated or regard discussions of this issue to be morally inappropriate.[48] In fact, the answer here isn't clear and we would very much like one to make progress on important practical questions, such as how valuable it is to save or avert lives.

8. Conclusion

This chapter began with the observation that many people hold what I called the Intuitive View—saving lives is good and, as the Earth is overpopulated, averting lives is also good—and that the Intuitive View is in internal tension. I expanded Greaves' work on this topic. I argued that, on Totalism, the Intuitive View is unlikely to be true and, if it is, neither saving lives nor averting lives would seem to be a high priority if we want to do the most good. I then argued that, on PAV, the Intuitive View is far more likely to be true and that, if it is, the values of saving lives and averting lives are in fact wide ranging. The obvious question to ask, given the importance of the effects of population size, is whether and to what extent the Earth is under- or overpopulated. I conclude, following Greaves and adding further reasons, that it's not obvious what the sign or magnitude of the other-regarding effects of extra human lives is. Further work is required here.

[48] (Greaves, no date) at section 4.

Chapter 3: Are you sure saving lives is the most good you can do?

0. Abstract

Many people believe that, if we want to do as much good as possible with our money, we should donate to effective charities that save lives. I set out four commonly-held but not exhaustive views of the value of creating and ending lives (such accounts are a combination of a population axiology with an account of the badness of death). For each view, I argue that, if we look just the self-regarding value of saving lives—and thus ignore the other-regarding impact of doing so—it is not obvious that saving lives is the best option. This non-obviousness results either from there being an alternative to saving lives that seems, at the first pass, to be more cost-effective, or because making comparisons is not straightforward. The overall conclusion is that, despite the intuitive appeal of saving lives, there are probably not many people who, if they reflected further on the empirical facts, would conclude, by the light of their own axiology, that it is the most good they can do.

1. Introduction

If we want to do as much good as possible with our money, one plausible thought is we should be using it to save lives. More specifically, the suggestion is that we should give to charities that provide cheap, effective health interventions in the developing world which primarily prevent the deaths of young children. I will use 'saving children' to refer to this specific method of preventing premature deaths, and 'saving lives', to denote to preventing premature deaths in general.[1] Singer and MacAskill emphasise the opportunity to save children, suggesting it is a leading

[1] To illustrate, an approach to saving lives that I do not have in mind is donating to charities that provide more swimming pool lifeguards.

altruistic option.[2] Singer even called a recent book *The Life You Can Save*.[3] For many years, GiveWell, a charity evaluator, claimed the Against Malaria Foundation ('AMF') was the world's top charity; AMF provides bednets that stop very young children (mostly under-5s) dying from malaria.[4] According to GiveWell, by giving to AMF, donors can save a (statistical) life for about $4,500.[5] GiveWell estimate that, of the roughly $100m or so of donor money they moved towards their top recommendations in 2017, $57m was spent on life-saving* charities.[6] Presumably many, if not most, of these donors choose to save children because they think it is where their money can have the greatest positive impact.

Furthermore, not only do many people seem to think it is true that saving lives (one way or another) is the most good you can do, they also seem to think this is *obviously* true. I have been surprised at how often, if I suggest something might be more valuable than saving lives, I encounter what David Lewis called 'the incredulous stare', followed by words to the effect, "Hold on. *Surely* it's better to save lives."[7]

The last two chapters serve to indicate saving lives is not obviously good, let alone obviously the most you can do, when we account for *other-regarding* effects of saving lives—the impact this has on everyone apart from the saved individual; specifically, the attention in the previous chapters was on the potentially negative impact that humans have on non-human animals (through creating unhappy

---

[2] (Singer, 2015) (MacAskill, 2015)
[3] (Singer, 2009) Admittedly, this may be a matter of marketing – Promoting the Good would have been a much less emotive title.
[4] In the period 2011 to 2016. See (GiveWell, no date b)
[5] See (GiveWell, 2019b), 'Nets' tab.
[6] (GiveWell, 2018b)
[7] (Lewis, 1986) This reaction is more plausible if we suppose it is capturing a normative, deontological intuition that we ought to save lives, even when we could bring about an alternative option that was more valuable (in terms of consequences).

animals for food) and on other humans in wider society (by putting pressure on available resources, i.e. what is normally referred to as a worry about 'overpopulation').[8] However, presumably those who do think saving lives is the most cost-effective option were focusing primarily on the *self-regarding* value of saving lives—the value solely related to the person whose life it is. They may also consider the loss the 'near and dear', such a parents, suffer when a child dies—which is an other-regarding effect—a reason to save lives too. I suppose this provides the far weaker reason to save lives: the loss to the child seems far smaller than the loss to everyone else combined (a point discussed in both of the previous chapters).

The next natural question to investigate, then, is whether saving lives—specifically, saving children—is the top beneficent option on what we might call the 'conventional analysis' of the issue; that is, if we just consider the self-regarding value of saving the children. As such, the conventional analysis leaves out all other-regarding effects. Investigating this question is of both theoretical interest and practical importance. Even if we thought accounting for various other-regarding considerations meant saving lives *could not* be the best option, it would still be intriguing and surprising if saving children turned out *not* to be the most good on the conventional analysis. However, as we saw in the previous chapter, the sign and magnitude of the other-regarding effects are not only unclear but hard to ascertain. Unless and until we are confident the other-regarding effects were large and negative enough to rule out saving lives as a possible top option. This question is still relevant for practical purposes.

---

[8] I note considerations of optimum population need not refer just to the impact humans have on other humans, although this seems to be what is normally referred to.

In this chapter, I set out what seems to be the four most commonly held views that one might take of the value of creating and saving lives, or what I'll call, for lack of a better term, a 'view of life-value'.[9] A view of life-value combines (1) a *population axiology*, a ranking of the value of states of affairs in terms of overall betterness (this is determined by specifying both (a) whose lifetime well-being matters and (b) how that lifetime well-being is to be aggregated) with (2) an *account of the badness of death*, a way of assigning lifetime well-being levels of individuals where the possible length of their life varies.[10] For each view of life-value—for brevity, I will usually just say 'view'—I argue it is not obvious that saving children is the best (i.e. most cost-effective) option. For three of the four views, this is because there is an alternative to saving children that seems, at first pass, at least to be roughly as cost-effective; for the remaining view, further specification is needed before it is straightforward to determine the cost-effectiveness. As noted, these comparisons ignore some of the other-regarding effects of saving lives, effects which are potentially large.[11] I make an exception when considering the fourth and final view, where it seems relevant to briefly assess how comparatively cost-effective it is to save children if one were to do so solely to prevent the grief that the family and friends would suffer from someone dying.

---

[9] I state the views later; stating them all first, and then discussing alternatives to saving children on them would make the paper less readable.

[10] It would be more accurate to call this an 'account of lifetime well-being', rather than an 'account of the badness of death': the way I've stated what such an account is prejudges certain issues in the literature on the 'badness of death', such as whether the badness of death is solely in how it affects lifetime well-being levels. For instance, (Kamm, 2019) outlines three factors that might make death bad besides the deprivation of future goods (and thus how the deprivation of such goods would reduce lifetime well-being). However, as the three accounts of the badness of death that I discuss here are standardly called 'accounts of the badness of death' in the literature, I use the more familiar (but less accurate) locution; see e.g. (Gamlund and Solberg, 2019) for a representative sample of this literature. I am here only interested in how death affects lifetime well-being as I am trying to connect the analysis of the badness of death with population axiologies (where lifetime well-being is standardly taken as the unit of aggregation).

[11] See chapter 2.

I suppose that the vast majority of people will think one of the four views, or at least something like them, is correct. The overall, surprising conclusion is this: despite the intuitive appeal of saving children, there are probably not many people who, by the lights of their own axiology, if they reflected further on the empirical facts, would conclude that it is the most good they can do. As such, individuals who do currently focus on saving children should consider if they can do more good by putting their resources towards something else.

This chapter is structured as follows. Sections 2 to 4 each introduce a view of the life-value of creating and ask if saving children is the best option. Section 5 concludes.

2. Totalist Deprivationism (TD)

The first view of the life-value combines the Totalism about population axiology with Deprivationism about the badness of death.[12] On Totalism, the value of a state of affairs is the sum of lifetime well-being of everyone who will ever live.[13] Importantly, Totalists hold creating a new life can be good/bad, and the value of doing is equal to the well-being of the person who is created. This contrasts with the other three views I discuss, which hold there is no value in creating new lives.[14] On Deprivationism, the badness of your death is the sum of well-being you would have had, had you lived.[15] Deprivationists will hold it's more valuable to save a 20-year-

---

[12] While, for simplicity, I refer only to Totalism, what I say this in section could, I think, be said of all population axiologies that give weight to all possible lives, i.e. Critical Level views, Variable value views and Averagism. For a summary of the different population axiologies see (Greaves, 2017)

[13] Population axiologies can (of course) include goods besides well-being. I follow the simplifying norm in the literature and restrict the analysis to well-being.

[14] Or, at least, no value in creating happy lives, those with net positive lifetime well-being. I return to this later.

[15] If more explanation is required: I take Deprivationism to hold that lifetime well-being is the unweighted sum of all the individual instances of momentary well-being (i.e. well-being at a time). As such, Deprivationism uses the same aggregation method to get from momentary well-being to

old than the 60-year-old as the former is deprived of more well-being than the latter, and this is more valuable by exactly the difference in the sums of well-being lost. I will (unimaginatively) call this Totalist Deprivationism ('TD')

It is not obvious, however, that saving children is the best way to do good on this view. One alternative is that, as TDs value all possible lives (not just the present generation), they should instead be putting their efforts towards preventing *existential risks*, that is, risks "that threate[n] the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development."[16] Beckstead summarises the steps in Bostrom's *Astronomical Waste* argument as follows:

1. The expected size of humanity's future influence is astronomically great.
2. If the expected size of humanity's future influence is astronomically great, then the expected value of the future is astronomically great.
3. If the expected value of the future is astronomically great, then what matters most is that we maximize humanity's long-term potential.
4. Some of our actions are expected to reduce existential risk in not-ridiculously-small ways.
5. If what matters most is that we maximize humanity's future potential and some of our actions are expected to reduce existential risk in not-ridiculously-small ways, what it is best to do is primarily determined by how our actions are expected to reduce existential risk.

---

lifetime well-being that Totalism uses to get from lifetime well-being to overall betterness. I note that (Bramble, 2018) argues there is no such thing as momentary well-being and lifetime well-being is the only genuine kind of well-being.

[16] (Bostrom, 2003) Not existential risks are not the same as extinction risks: the former are a wider category that includes concern for non-extinction outcomes that could reduce 'future development'.

6. Therefore, what it is best to do is primarily determined by how our actions are expected to reduce existential risk. [17]

To make the expected value of the future plain, here's a quote from Bostrom:

> Suppose that about 10^10 biological humans could be sustained around an average star. Then the Virgo Supercluster could contain 10^23 biological humans. This corresponds to a loss of potential equal to about 10^14 potential human lives per second of delayed colonization.[18]

Bostrom's particular concern is the threat that emerging technologies, such as artificial intelligence, pose to sentient life.[19] As such, so the thinking goes, those who care about all possible lives should focus their efforts on reducing existential risks, rather than anywhere else, and that includes saving children.

To make the scenario both more concrete and less reliant on lives in the very far future, let's look at the value of saving humanity if it lives only 100,000 years longer (a fraction of the time it could exist for). Assuming there are 7 billion people per century, then means 7 trillion people's lives are at stake. Suppose there is a 1% risk of extinction over this century (i.e. the risk this year is 0.0001, the same the next, etc.), and none thereafter.[20] Assume that there is some project that would cost $1bn per year, such as an asteroid tracking and deflection programme, and this would reduce the risk proportionally by 1%, thus, if we spent a billion each year this century, the cumulative probability of extinction declines from 1% to 0.99%.[21] If we spent $1bn this year, that would lead to the existence of 70 million more lives in

---

[17] (Beckstead, 2013)
[18] (Bostrom, 2003)
[19] (Bostrom, 2014)
[20] This model is partially adapted from one proposed by (Lewis, 2018)
[21] (Beckstead, 2013) ch.3 argues asteroid deflection would be more valuable that 'proximate benefits' e.g. saving children, but does not offer a quantified comparison.

expectation. Assuming that the expected value of financial contributions is linear, then, if we scale the numbers down, $10,000 put towards an existential risk reduction organisation would create 700 lives, in expectation.

According to GiveWell, the Against Malaria Foundation saves, in expectation, slightly more than 2 lives of under-5-year-olds for $10,000.[22] Let's assume for simplicity this save two lives and saving these lives is as good, in terms of the individual value related to those lives, as creating 2 new people that live their entire lives. Assuming, again for simplicity, all human lives now and in the future have the same lifetime well-being, then preventing existential risks—'saving humanity'—seems to be over 300 times more cost-effective than saving children.

At the first pass, then, saving children does not seem the most cost-effective option on this view. That suffices for our purposes, hence a longer discussion is outside the scope of this analysis.[23]

3. Person-affecting deprivationism

In this section, I assess the importance of saving children on a view that combines a *Person-Affecting View* about population axiology with Deprivationism (about the badness of death). Call this combined view *Person-Affecting Deprivationism*

[22] (GiveWell, 2019a)

[23] Briefly though, I note that advocates of saving children would need to take issue with premises (1) or (4) from Beckstead's summary. If the value of the future is very small – say, we expect future lifetime well-being to be on average zero after this generation – or we think we are powerless to change the value of the future, the, we should focus on doing good in the near-time, i.e. the lifetimes of humans alive today; saving children may well be the best candidate here, although this still isn't obvious, as I discuss in the next sections. If we can make substantial changes in the value of the future, it seems Totalists will think they should do that, even if the expected value of the future looks very bad. A further option would be to suggest we should revise our use of expected value in situations where there are low probabilities of very high stakes effects, so-called 'Pascal's Mugging' scenarios. For discussion of these, see e.g. (Bostrom, 2009)

('PAD'). I'll explain what this view is and then compare the cost-effectiveness of saving children to saving (non-human) animals.

Person-Affecting Views that hold, when evaluating states of affairs in terms of their overall betterness, a subset of all possible people are deemed 'extra', in the sense that their welfare does not matter (or matters for less) when assessing overall betterness. Different Person-Affecting Views are characterised by who counts as 'extra'.[24] I will drop the inverted commas from 'extra' hereafter.

Prominent examples of Person-Affecting Views are Presentism (only presently existing matter), Actualism (only actually existing, as opposed to merely possible, people matter), and Necessitarianism (the only people who matter when deciding between outcomes are those who exist whatever we choose to do). Each of those three views, as stated, has the important implication that there is no value in creating new lives: a hypothetical person we are considering creating does not presently exist, will not actually exist (if they are not created), and does not (given the choice context) necessarily exist.

The following analysis is not sensitive to a choice between the three examples above, so I will simply use talk of a 'Person-Affecting View(/s)' without specifying any of three.[25]

---

[24] This conception of Person-Affecting Views and the use of the word 'extra' is taken from (Beckstead, 2013).

[25] The 'Harm-minimisation' theories are another type of Person-Affecting View. The basic idea is that, when considering different possible states of affairs, the amount of (comparative) harm an individual suffers in a given state of affairs is equal to how much worse off they are in that state than the one where they had maximum well-being. The best state of affairs is the one where there is the lowest comparative harm, which is summed over all the people who exist in that state of affairs. Such views function along the lines of the asymmetric Person-Affecting View that I mention momentarily in the text. For simplicity, I do not discuss them separately. See (Greaves, 2017) at section 5.3 for a brief, (critical) discussion and further references.

Person-Affecting Views can be as *asymmetric*, on which people who have lives that would not be worth living are not considered extra. *Symmetric* Person-Affecting Views hold that whether a life is living or not does not affect whether it is extra or not.[26]

We need to attach a theory of the badness of death to our (generic) Person-Affecting View. In this section, we combine a Person-Affecting View with Deprivationism about the badness of death to get Person-Affecting Deprivationism (PAD). In the next two sections, we combine Person-Affecting Views with different accounts of the badness of death, which I will explain when we get there.

What might be better than donating to saving children on PAD? I consider two options (1) reducing extinction risks, (2) saving animals. These are, of course, not the only alternatives.

---

[26] There are further distinctions to draw. The following two seem standard and are used by e.g. (Beckstead, 2013) at p. 75. First, between views that are strict (extra people do not matter) versus moderate (extra people carry some weight, but less than those who 'matter' fully). Second, between narrow vs wide versions of Person-Affecting Views, where on the latter, extra people are not discounted. This standard distinction doesn't seem right; there seems to be no relevant distinction between strict and narrow views – in both, extra people do not matter. Hence, the distinction should be between strict/narrow views, moderate views, and wide/(loose?) views, where the views can be understood as points on a spectrum of how much extra people matter, which ranges from 'not at all' to 'fully'.
The other issue is we have not drawn a distinction between views that hold, when discussing which people matter, and whether 'people' refers to the people *de re* or the people *de dicto*. This is crucial for Necessitarianism. For a '*de re* Necessitarian', the people who matter are the specific individuals who will exist whatever choice is made. In effect, for reasons discussed later in section 3.2, only present people matter. For a '*de dicto* Necessitarian, the people who matter are those who will exist, whoever they happen to be. While *de dicto* Necessitarians will not think creating new people matters, they will want to increase the well-being of the future people, *whomsoever* they are, and will think that what matters is what happens to future generations. Importantly, a *de dicto* Necessitarian is different from a Wide (version) Necessitarian – only the latter of the two holds that creating new people matters. I am not aware this distinction has been made before. Note that the *de re*/*de dicto* distinction makes little sense on Presentism (the *de dicto* and *de re* present people as the same set of individuals) or Actualism (the actual people *de dicto* are the same as the possible people). To clarify, for simplicity, I will be using strict, *de re* Person-Affecting Views. (Bader, no date) advances what I call a *de dicto* Necessitarian view and he calls a 'Same-Number' Person-Affecting View.

3.1 Existential risks

Given reducing extinction seemed more cost-effective than saving children on TD, a natural thought is whether this is also the case for PAD. The argument for the importance of extinction risk reduction on TD was based on the huge value that results in future generations existing. On Person-Affecting Views, what happens to these future generations is of no concern—they consist in non-present, (presumably) non-necessary and, if they do not exist, non-actual people.[27] However, on PAD, it would be bad if everyone *died now* from some catastrophe, as that would deprive the current generation of the goods of life. Is extinction risk reduction a plausibly cost-effective alternative to saving children on PAD?

I am only aware of one attempt to crunch the numbers on this, which is by Gregory Lewis.[28] Lewis' model takes as inputs the fact there are 7.6 billion people on Earth, the worldwide mean age is 38 and worldwide life expectancy is 70.5. Thus, he states the 'naïve' loss if everyone died tomorrow would be 32.5 years per person on average, meaning the total loss is 247 billion life-years. He uses the same numbers as we had before regarding the risk of extinction (1% this century, uniform by year) and the tractability of existential risk reduction ($1bn a year over a century reduces the existential risk from 1% to 0.99%).[29] On his model, the 'cost per life-year' is $9,200.[30]

---

[27] At least, this is the case on strict de re Person-Affecting View. For instance, wide Person-Affecting View will care about future generations. See 26 for further explanation.
[28] (Lewis, 2018)
[29] These are identical as I used Lewis' numbers for my earlier estimate.
[30] In fact, matters are more complicated for a theoretically interesting but practically unimportant reason. Lewis models the cost-effectiveness of the intervention as being the same each year, i.e. 1 year from now, 10 years from now, etc. However, on Person-Affecting Views, only a subset of all possible matter: the present, necessary, or actual people. The further away an event occurs *from now*, the fewer of these people will exist; hence, the further in the future in event occurs, the less valuable it will be, *ceteris paribus*, as the 'pool' of people who matter shrinks. This is easiest to see

By comparison, AMF is estimated to save an under 5-year-old child's life for about $4,500.[31] Assuming such a child is 2.5 and would live a further 60 years, the 'cost per life-year' for AMF is $75.

Lewis concludes that while existential risk reduction "compares unfavourably to best global health interventions, it is still a *good buy*: it compares favourably to marginal cost effectiveness for rich country healthcare spending" (emphasis in the original).[32] This seems sufficient to say that saving children is obviously better than reducing extinction risks on PAD.

3.2 Saving animals

A more promising alternative to saving children on PAD is 'saving' animals.[33] Most of those sympathetic to Person-Affecting Views opt for the asymmetric version. On the asymmetric version, preventing the existence of bad lives lived by non-animals will be valuable. Plausibly, the great majority of animals in factory farms live bad lives. Arguably, these animals could be helped very cheaply.

The Human League runs campaigns for individuals and organisation adopt behaviours that reduce farmed animal suffering. According to Animal Charity Evaluators (ACE), The Human League spares 0.56 farmed animal life-years per

---

with Presentism. Suppose money I spend now prevents an asteroid strike in 1,000 years. Given none of the people who presently exist now will exist then, this has no value on presentism. Suppose money I spend now prevents an asteroid strike in 30 years. Presentism only values the effect this has on the people under who will *then* be age 30 or old. And so on. We can see how Person-Affecting Views imply what we can call *contingent time discounting* ('CTD'). CTD functions a bit like *pure time discounting* ('PTD'), idea that we should reduce the value of future events by N% for every year it is further in the future, solely on the basis it is further in the future. I call CTD 'contingent' as the discount occurs *over* time but is not justified *merely on the grounds time has passed*. This isn't practically important here saving children is already many time more cost-effective than preventing extinction on PAD. I intend to explore CTD in future work.

[31] (GiveWell, 2019b) set 'Nets' table. The precise figure is $4,388.

[32] (Lewis, 2018)

[33] The scare quotes reference the fact it's questionable whether preventing something from ever existing counts as 'saving' it.

dollar.[34] These estimates are admittedly uncertain—ACE's 90% subjective confidence intervals are that $1 spares between -0.84 and 6.9 life-years per dollar. The estimates are also quite complicated, more complicated than it would be useful to unpack them here. Nevertheless, suppose we are prepared to take the figures at face value, then the same size donation to the Human League spares about 150 animal years for every 1 human year that AMF saves (i.e. $0.5 vs $75 per life-year). Without wishing to repeat the analysis in chapter one, assuming the humans live happy lives and the farmed animals live unhappy ones, unless one thinks the human's positive well-being is 120 times greater in magnitude than the farmed animal's negative well-being, the saving of animals looks to be more cost-effective.[35] There is at least a case that money spent on the animals does more good than saving children.

What if one held a symmetric Person-Affecting View? On this, roughly speaking, the well-being of future entities, human or non-human, does not count. Why would this be the case? Suppose one donated to the Human League's campaigns. Given the short life-span of many farmed animals—broiler chickens live only seven weeks—your money would not plausibly increase the welfare of the specific animals that presently exist. A successful campaign would (presumably) cause different particular animals to be created, hence you do not help any that exist necessarily.[36] That accounts for presentism and necessitarianism. According to actualism, the betterness relation between two alternatives depends on which outcome is actual;

---

[34] (Animal Charity Evaluators, 2018b) at footnote 84 provide a link to ACE's Guesstimate (cost-effectiveness) model of The Humane League.
[35] Or one values goods besides well-being and/or applies a pure-species and the combined change the ordering of cost-effectiveness. See the 'discounting' objections in chapter 1.2.1.
[36] This would not be the case on a *de dicto* Necessitarian view. See footnote 26 for an explanation.

here, as you are choosing which lives become actual, actualism provides no guidance with respect to betterness.[37] As such, on a symmetric Person-Affecting View, funding efforts to spare factory-farmed animals does not look valuable.

Arguably, however, on a symmetric Person-Affecting View, it would be better to donate to animal shelters rather than save children. The symmetric PAD will hold that animals which currently exist matter.[38] Donating to animal shelters to help, e.g. cats and dogs find homes instead of being put down, will count as valuable and, given Deprivationism, will hold that the value is the sum of well-being the animals would have if they lived.

Now for some numbers. ACE estimates that $1,000 given to animal shelters saves seven animal lives.[39] Let's suppose the animals in question are dogs and cats and would live 8 more years if they were re-homed. For $3,500 then, it seems you'd saved 196 years of animal life. Compare this to AMF, which for $4,500, we supposed saved 60 years of human life. Echoing the analysis about farmed animals, someone who thought that cats/dogs had levels of well-being not much lower than that of humans could reach the conclusion it's better to donate to animal shelters. [40] This analysis is interesting because many animal advocates argue that, if you want to help

---

[37] The actualism violates the condition of 'axiological invariance' condition: that the betterness relations between two alternatives should not depend on which alternative is actual. This is indeed puzzling and many philosophers have taken this as being a decisive reason to reject actualism. For discussion of this (and the related condition of 'normative invariance') see (Broome, 2004) p.74 and (Bykvist, 2008)

[38] Present lives are also necessarily both necessary and actual lives.

[39] According to (Animal Charity Evaluators, 2018a) this is the cost for an animal shelter to rescue a cat or a dog. It's not clear what the counterfactual of these costs not being met is—would the animal die on the streets, be put down, or perhaps something? I looked at the website of several animal shelters they all claimed they did not put animals down. This case may be merely hypothetical ("If one could donate to an animal shelter which would otherwise destroy the animal, then …") but it is nevertheless an *interesting* hypothetical.

[40] The dogs I've known seem happier than the average human, but not the cats. I note that proponents of Mill's higher/lower pleasures distinguish may well think it is "better to be Socrates dissatisfied than a pig satisfied" but it is beyond the scope of this paper to discuss this topic. (Mill, 1861)

animals, you should focus on factory farming rather than giving to animal shelters (in reality, the latter gets enormously more resources).[41] However, on this particular view, animal shelters would be more cost-effective than factory farming, and maybe even than saving children.

4. The Person-Affecting Time-Relative Interest Account (PATRIA)

The third view of life-value is also a Person-Affecting View but swaps a Deprivationist account of the badness of death for the Time-Relative Interest Account (TRIA). I will refer to this combined view as 'PATRIA'. I'll motivate and explain the Time-Relative Interest Account and say why making cost-effectiveness comparisons on this view is problematically arbitrary.

On TRIA, the badness of death is a function of (a) the well-being that a person would have had if they'd lived (as with Deprivationism) and (b) the strength of the psychological connection that person has to their later self.[42] Whilst Deprivationism and TRIA agree it's better to save a 20-year old than an 80-year old, they disagree on whether it's better to save a foetus or a 20-year old. On TRIA, it's better to save the 20-year old, despite the fact the foetus will (all else being equal) have 20 years more life to live, as that foetus will be greatly psychologically different from its later self.[43] As Nils Holtug explains:

> After all, foetuses and infants usually have rather simple psychologies and thus few of the preferences, memories and character traits they will acquire later in life. Assuming an appropriately large discount rate, then, the Time-

[41] (Animal Charity Evaluators, 2016)
[42] See (Liao, 2007; McMahan, 2002; McMahan, 2015)
[43] It is an open question as to exactly which of the psychological relations matter: it could be memory, personality, or something else.

> relative Interest Account implies that the 20-year-old will actually have a stronger interest in survival than the infant or foetus has.[44]

There are questions both about how exactly this view is formulated, and whether it should be understood as a view of the badness of death. As Greaves states, when deciding how valuable saving a life is on TRIA, we need to know:

> *precisely which person-stages* [i.e. the moments of a person's life] *count*? Are the relevant time-relative interests, for instance, only those of present person-stages ('presentism')? All actual person-stages ('actualism')? All person-stages who will exist regardless of how one resolves one's decision ('necessitarianism')? All person-stages who would exist given *some* resolution of one's decision ('possibilism')? Or something else again? (emphasis in original).[45]

Greaves goes on to argue that both the present and actualist person-stage versions of TRIA are implausible (for reasons that need not detain us here) and suggests TRIA is not capturing the axiological badness of death, i.e. how bad it is in terms of final value, but rather our emotional reaction to how bad it seems when someone dies at different ages. This strikes me as correct.

However, even if this view is somehow confused, it is still useful to try to understand the implications it would have on the question to hand. Many people, when thinking about the value of saving children, discount the value of saving children's lives relative to those of adults, because of TRIA-like reasoning. GiveWell, a charity evaluator, explicitly includes a discount for saving those under-5 vs over-5 years old

---

[44] (Holtug, 2011)
[45] (Greaves, 2019)

in their cost-effectiveness model in.[46] So we will push on. For our purposes, it is sufficient to use the present-stage version of the view.

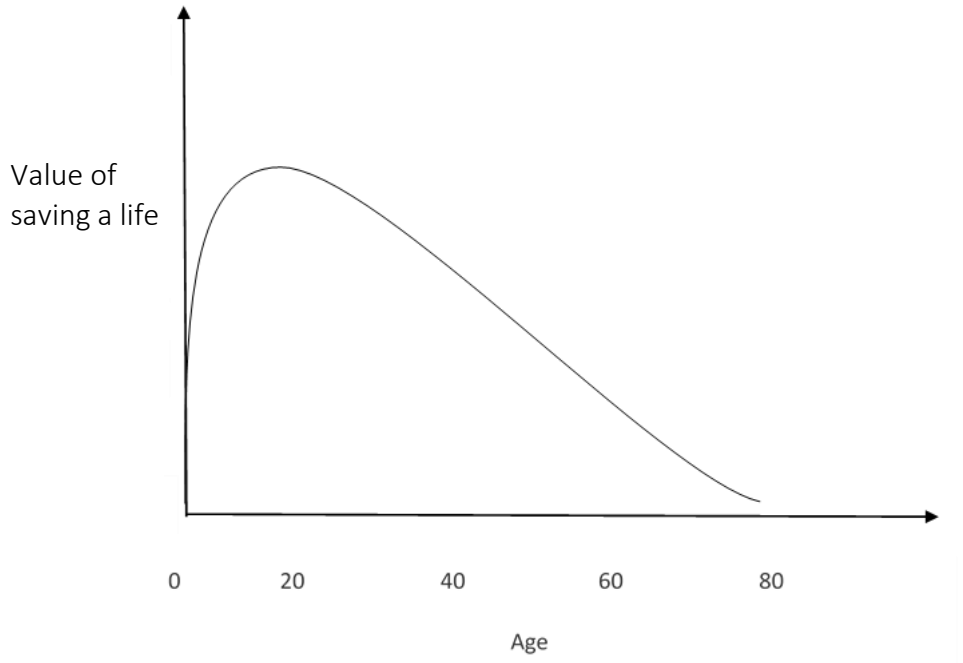

Figure 3.1. Value on TRIA of saving lives at different ages

Schematically then, on TRIA, the value of saving lives that are presently of different ages looks something like the curve that is represented in figure 3.1. It is more important to save those in their 20s/30s. Younger people will not be so psychologically connected to their later selves; older people will have relatively greater psychological stability but have less time left to live.

Two observations. First, as is already salient, TRIA discounts the value of saving children—they have a lesser psychological relation to their later selves than mature people do. Hence, saving children is far less valuable on TRIA that it is on the

[46] See (GiveWell, 2019a) at the 'Moral Weights' tab. Different GiveWell staff members have confirmed (personal conversation) that members of their organisation value saving under-5s less than over-5s as a result of TRIA-type concerns as well as the larger other-regarding effects of older deaths, e.g. the greater grief felt by parents.

Deprivationist view.[47] Of course, thoughtful TRIA-advocates who choose to save children will already have incorporated this; I am not claiming anything surprising here.

Second, while we know saving 20-year-olds is better than saving 2-year-olds on TRIA (as that is the intuition the view is meant to capture), it's unclear precisely *how much better* the former is than the latter. 10% better? Twice as good? 10 times as good? There seems to be plenty of room for disagreement about what the shape of the curve should be, and such questions need to be (somehow) settled before the TRIA-advocate could work how to do the most good.[48]

In the previous section on PAD, we compared saving animals to saving children and said the former looked perhaps to be competitive with the latter. The natural question to ask is whether this is still the case on PATRIA. As saving children is less valuable than saving adults on PATRIA, we might think that saving animals is *even better* on PATRIA than on PAD. In fact, matters are more complicated. As McMahan observes, many people think, for TRIA-like reasons, that while preventing animal suffering is important, altering the lengths of animals is not important because:

---

[47] Hilary Greaves raises the question of whether such intertheoretic comparisons make sense, not intertheoretic comparison are deemed problematic in the moral uncertainty literature – e.g. see (Bykvist, 2017) There doesn't seem a problem here: to determine the badness of death, both Deprivationist and TRIA take as inputs the (sum of momentary) well-being the person doesn't experience by dying early. The only difference is TRIA additional applies a psychological discounting before determining the axiological badness of the death. Hence, it's easier to see how, if one switched from being a Deprivationist to being a TRIA advocate, it would be a simple matter to rescale the value of different outcomes.

[48] For examples of such a disagreement, see (Norheim, 2019) who argues an 'extreme' version of TRIA. which holds we should save adults over children is untenable, although a more a 'moderate' one is acceptable. (McMahan, 2019), writing in the same volume, accepts the 'extreme' version. A further issue is that, as TRIA discounts the value of future well-being by how psychologically connected the present self is to that later person, this we also need the later lives of those who have reached adulthood: adults are not psychologically identical to their later selves either, whatever account of psychological connection we use (personality, memories, etc.). Hence, TRIA requires a further discount. As accounting for this complicates matters without changes the result, I leave it to the side.

> They [i.e. animals] are not self-conscious, or are self-conscious only to a rudimentary degree, they are incapable of contemplating or caring about any-thing more than the immediate future. They do not, therefore, have desires or intentions or ambitions for the future that would be frustrated by death.[49]

What seems to follow is that if one had an asymmetric Person-Affecting View and combined that with TRIA, sparing animals from living in factory farms would look relatively *more* cost-effective, compared to saving children, than it did on the asymmetric Person-Affecting View combined with Deprivationism. This is because saving children gets discounted but sparing bad animal lives does not.

Matters are (even) less clear on a *symmetric* Person-Affecting View that is combined with TRIA. This view will want to discount donating to animal shelters—there is little value in extending the lives of presently existing animals. As both donating to animal shelters and saving children are discounted, it is not clear, without specifying further details of the view (and empirical matters), which is supposed to be more cost-effective.

The reader may wonder why I have not attempted a more obvious comparison between saving children and an alternative that improves lives by increasing the well-being of individuals whilst they are still alive.[50] In chapter 7, I compare saving children against treating mental health. I do so assuming Deprivationism and using subjective well-being (SWB) scores, where individuals say how satisfied they are with their life on a 0-10 scale. The issue with the comparison, as I discuss there

---

[49] (McMahan, 2008)

[50] I note I am using 'improves lives' here differently from how I standardly use it in the thesis. On my typical usage, improving lives refers to increasing happiness during a life.

(chapter 7.3.2) is whether the life-saving or life-improving interventions are more cost-effective (as measured by their subjective well-being impact). This is highly sensitive to a currently arbitrary decisions about where to place the 'neutral point' equivalent to non-existence on the 0-10 scale. If it's at 5 (it's hard to believe it could be any higher) then treating mental health is more cost-effective.[51] If the neutral point is 0, saving children is perhaps seven times more cost-effective. This analysis is in terms of Deprivationism, and saving children is less valuable on TRIA. It's not clear saving children is seven times less valuable on TRIA, and hence the saving children-treating mental health cost-effectiveness comparison still turns on where the neutral point is.

I propose to leave matters here. I take the foregoing analysis to be sufficient to show it is not obvious whether saving children is the best option on PATRIA-type views.

5. Person-Affecting Epicureanism

The fourth and final view we can momentarily consider conjoins a Person-Affecting View with Epicureanism about the badness of death—we can call this 'Person-Affecting Epicureanism'. Named after the ancient Greek philosopher Epicurus, Epicureanism is the position that death is not better, worse, or equally as good as living is for a person, regardless of how much well-being they would have had if they'd lived; there is no value in living a longer (or shorter) life.[52] As such, what Epicureans will say about adding person-stages is analogous to what Person-Affecting Views will say about adding lives to a population, which is that, in each case, doing so does not matter in the appropriate sense—for the former, adding

[51] In fact, saving children would then be bad, given that average life satisfaction in target countries is sometimes below 5: see (Helliwell, Layard and Sachs, 2018).
[52] (Epicurus, 2019)

person-stages does not contribute to lifetime well-being levels, for the latter, adding lives does not contribute to the overall betterness of a state of affairs.

Without getting into the possible mechanics of view, it's clear that Person-Affecting Epicureans will not think saving lives is a promising way to do good if they are doing the 'conventional analysis' of just considering the self-regarding value of saving lives.[53] Of course, Epicureans will still hold that dying, rather than being dead, can be bad for me, as dying happens while I am still alive; and they will allow that my death can be bad for others. Epicureans will more naturally focus on improving lives.

One might wonder if saving children could be the most cost-effective option for Person-Affecting Epicureanism, solely because it prevents bereavement-based suffering among the living. I propose to quickly check this with figures that I mostly calculate and explaine elsewhere in the thesis. In footnote 23 in chapter 2.3, I estimate the loss of a child is *at most*, 4 life satisfaction point-years ('LSPs'), where 1 LSP is equivalent to an increase of one person's self-reported life satisfaction by 1 point for a year on a 0-10 scale. (I argue in chapter 4 that self-reported life satisfaction scores are a reasonably proxy for happiness). Averting this loss through AMF would cost around $4,500. Hence the cost-effectiveness is 1.1 LSPs/$1,000. In chapter 7.2, I estimate that the provision of mental health treatment via StrongMinds, a charity working in low-income countries, is 7.4 LSPs/$1,000. Hence, saving children for the sake of averting bereavement-grief seems less cost-

[53] Epicureans seem to have analogous choice points about which person-stage matter as Person-Affecting Views have about which persons matter – the present, necessary, or actual person-stages/persons. The (psychologically) consistent thing for Person-Affecting Epicureans to do would be to make the same choice at both places, e.g. only necessary people matter and, what's more, only the necessary person-stages of those people matter.

effective than simply treating unhappy people directly via improving their mental health.

Hence, (unsurprisingly) saving children does not seem to be the most cost-effective option on Person-Affecting Epicureanism either.

6. Conclusion

I've specified four views of life-value and suggested it is not obvious that saving children is the most good you can do on any of them, either because something else seems more cost-effective or it's not straightforward to assess the cost-effectiveness of an alternative. I note I was only examining the self-regarding value of saving children–apart from in section 3.5–which was a deliberate limitation of the analysis. This survey does not cover all possible views of life-value, but I suspect the vast majority of people will think these views—or ones very much like them—are correct. It's possible there is some view of life-view I have not considered on which saving children would be the highest impact option, but I cannot think what it would be. What follows is that, for most people, saving children is not (or at least, not obviously) the most good they can do by the light of their own views of life-value. Hence, if they currently put their resources towards that aim, they should consider putting it towards somewhere else that would allow them to do even more good.

Chapter 4: Happiness for moral philosophers

0. Abstract

The last few decades have seen an explosion of research in the social sciences into 'subjective well-being' (SWB), self-reported measures of happiness and life satisfaction. Moral philosophers seem not to have taken much notice of this. Although MacAskill and Singer seem to both to hold well-being consists in happiness and draw on SWB studies in their 2015 books on effective altruism, they ultimately justify their charity and career recommendations by appealing to more conventional proxies for well-being— income and health metrics, such as Quality-Adjusted Life Years (QALYs). I suggest four possible objections to using self-reports, rather than of any other metric(s), to determine what increases happiness: (1) happiness can't be measured through self-reports; (2) individuals' happiness scores cannot be meaningfully compared; (3) there isn't yet enough data on self-reported happiness to guide our decision-making; (4) using self-reported measures is unnecessary as it wouldn't change our priorities. In this chapter, I meet the first objection fully and the latter three partially.

1. Introduction

I start with some preliminary general remarks, whose relevance will shortly become clear. I think these are—or should be—uncontroversial.

When we disagree about how to do the most good, the source of this disagreement could be over which axiology is correct—an *axiology* is a method of ranking

outcomes in terms of their final value.[1] Alternatively, it could be over the (non-evaluative) facts—our understanding of how the world is and how it will be. Of course, we could disagree about both. Importantly, once we've settled on which axiology we're using, we can only being disagree about the facts. For instance, imagine two classical utilitarians disagree about whether it would be better to do A or B. They agree the value of an outcome is the unweighted sum of happiness in it. Suppose, further, they agree that happiness is defined in a particular way. As they entirely agree on matters of value, they can only be disagreeing about matters of fact.

Suppose we haven't measured the quantities of happiness in A or B. In this case, our assessment of which has more is a *subjective judgement of fact*, and not (as we might have suspected) an axiological judgement. The situation is analogous to the one where we're trying to guess whether giraffes or buffaloes weigh more—it is just a (tricky) subjective judgement of the facts.

It might seem odd to say that the amount of happiness there is in different outcomes is a factual claim, rather an axiological one. To see that it is merely a factual claim, note that we can try to assess how much happiness there is in various outcomes while at the same time holding that happiness has no value at all. There's nothing

1 We could disagree over the component parts of an axiology, namely, (a) which thing of things are intrinsically valuable and constitute the good(s) (e.g. well-being, equality, justice, etc.), (b) how to aggregate the goods to determine the overall value of a state of affairs (unweighted sum, priority-weight sum, 'maximin', etc.), and (c) who the value-bearers of these goods are (e.g. all possible people, present people, necessary people, etc.). The first two components are individually necessary and jointly sufficient for a fixed-population axiology. That is ranking outcomes where the number of people is invariant. The addition of the third component yields for a variable-population axiology, which is needed where the number of people is not invariant, e.g. when considering the value of new lives.

special about happiness here—the same would apply whichever possible intrinsic good was being considered.[2]

Moral philosophers, I take it, have no special powers for judging facts.

If we are disagreeing about matters of fact and we can measure (or test) those facts, then we should take that measurement (or test) as authoritative over our subjective judgements. If we could weigh the buffalo and giraffe on a set of working scales, we would, I presume, defer to the scales.

We now turn to the subject of this chapter, which is happiness and its measurement.

While there is a long-running suspicion, primarily arising from economics, that happiness cannot be measured, this hasn't stopped some social scientists from trying. Happiness research, sometimes called 'subjective well-being' (SWB) research, terms I distinguish shortly, is now a major enterprise with about 170,000 books and articles published in the last 15 years.[3] Policymakers are starting to take note: the UK government has been measuring SWB since 2010; the OECD, an economic organisation comprising mostly wealthy countries, suggested in 2013 that its member-states should do likewise.[4] This increasing interest is driven, at least in part, by the belief among advocates that measuring happiness is serious science.

While philosophers generally believe that increasing happiness is good (intrinsically or instrumentally), they have taken little interest in this burgeoning field of happiness research so far. A few philosophers of science have written a few pieces

---

[2] Assuming the good can be quantified.
[3] (Diener, Lucas and Oishi, 2018)
[4] (ONS, 2018), (OECD, 2013)

on whether happiness can be measured.[5] The response from moral philosophers has been even more muted.[6] In their recent books on effective altruism, William MacAskill and Peter Singer claim (or imply) that happiness is the only intrinsic good and draw on SWB studies.[7] However, both primarily justify their suggestion that global poverty is the priority by appealing to more established proxies for happiness, such as standardised health metrics, e.g. Quality-Adjusted Life-Years (QALYs) and income.[8] Their particular charitable recommendations are largely drawn from the research of GiveWell, a charity evaluator, which also largely relies on the same, more conventional metrics (as opposed to research on SWB).

In light of what has been said so far, it should be clear that the question of what increases happiness is factual, once axiological matters are settled. If it were solely an axiological question, philosophers could (perhaps) safely ignore the social scientists' activities.[9] As it is, if the social scientists are correct about the facts, then moral philosophers, insofar as they want to make recommendations about how to increase happiness, should base their recommendations on that evidence, assuming it is practically feasible. Of course, if the social scientists are mistaken, then their efforts to collect and use such evidence are misguided (and someone should let them know).

---

[5] For philosophers of science, see e.g. (Alexandrova, 2017), (Haybron, 2016), (Feldman, 2010b), (Tiberius, 2006).
[6] It's unclear if moral philosophers have written on this topic at all.
[7] (Singer, 2015) p98 writes "something is a sacrifice if it causes you to have a lower level of well-being or, in a word, be less happy". (MacAskill, 2015) p45 seems to equate well-being with subjective well-being, of which happiness a component. I return to subject well-being in section 2.
[8] For use of happiness studies, see (Singer, 2015) p98 and (MacAskill, 2015) p27. MacAskill, op. cit. p44 says he will mainly use QALYs as a metric for well-being. Singer, op. cit. pp. 130 uses QALYs and Disability-Adjusted Life Years (DALYs) to make various trade-offs.
[9] (Angner, 2013a) argues Fred Feldman holds the view that empirical research is not relevant to philosophical conclusions and that Feldman is both mistaken and an outlier in this regard.

What objections might someone raise to relying solely (or, at least, primarily) on the subjective well-being research to determine what would increase happiness? There seem to be four.

First, the measures of happiness used in social science are not *valid*, that is, they do not succeed in capturing the underlying phenomenon they aim to capture. A slightly weaker version of this objection is that they are valid only sometimes and hence cannot be fully relied on.

Second, individuals' self-reported happiness score cannot be meaningfully compared or aggregated. Hence, we can't use self-reports to tell us what increases the sum of happiness and have to turn instead to other methods. More technically, the concern is about whether the scales are *interpersonally cardinal*—that a one-point increase on a (say) ten-point scale for one person represents the same increase in SWB for anyone else.

Third, there isn't enough available evidence on happiness to work out what the practical implications are even if we wanted to, hence we must rely on other proxies.

Fourth, it's unnecessary to start using the empirical happiness research because it wouldn't change our priorities anyway.

Neither Singer nor MacAskill explain why they draw on SWB data but don't ultimately base their suggestions on them. Hence, it's unclear which of the objections they find plausible. I suspect the truth of the matter—and the most charitable conclusion, is that some combination of the four is at play. Angner

suggests the objection among philosophers more broadly is that philosophers do not think happiness is measurable.[10]

In this chapter, I aim to meet the first objection entirely the latter three partially.

Here is the prospectus. Section 2 sets out the relationship between happiness and SWB and proposes, in light of the comparative scarcity of evidence on more theoretical ideal measures of happiness, that measures of life satisfaction are a suitable proxy for happiness—this helps respond to the third objection.

Section 3 and 4 draw us into the philosophy of science. Section 3 argues that the social scientists have been getting it right—they are succeeding in measuring happiness. I explain and defend the background theory in the philosophy of science, 'construct validation', on which SWB can, *in theory*, be measured. I then present the evidence demonstrating that current SWB measures are, *in fact*, 'valid', that is they succeed in measuring what they aim to measure. My reasoning here is not original. I will be restating arguments made by philosophers of science and social scientists. However, these arguments are usually made in isolation: philosophers tend to say why happiness could, in theory, be measured without offering the evidence that supports the claim our current measures are, in fact, satisfactory;[11] social scientists tend to do the reverse, offering the supporting facts without fully explaining their relevance.[12] My aim is to conjoin these arguments to give a comprehensive, but brief, account of why happiness can be and is being measured.

---

[10] (Angner, 2013b)
[11] Examples of this see (Angner, 2013b) and (Alexandrova, 2016)
[12] And for examples see (Dolan and White, 2007) and (Diener, Inglehart and Tay, 2013)

Section 4 addresses the concern about interpersonal cardinality of the self-reports; this topic does not seem to have attracted much discussion. I identify six conditions which are jointly sufficient for interpreting the 'raw' SWB scores as 'universally' interpersonally cardinal—I explain the words in inverted commas in the section itself. I then offer an initial assessment of each condition. Some of these seem highly plausible. For others, we cannot put all our doubts to rest and I identify what further work is needed. I follow this up by arguing that, if the raw scores are not cardinal, we can make them cardinal by applying the appropriate mathematical transformation. Hence, if there is a lack of cardinality this is not, in principle, a problem. I propose we should assume the raw scores are cardinal unless and until new evidence suggests they are not—they seem at least roughly cardinal and there doesn't seem to be a particular transformation we could apply that would take the raw scores closer to cardinality.

Section 5 takes the initial steps required to reply to the third and fourth objections and launches us from the philosophy of science into applied ethics. I observe that, if we look at the evidence on happiness, mental health suddenly stands out as a major problem; one not mentioned by Singer or MacAskill or, indeed, effective altruists more broadly. I suggest this result is somewhat unsurprising when we consider how QALYs underweight the badness of mental illness. That a potentially major problem seems to have been missed seems good cause to re-evaluate what our priorities are if we want to do the most good and the means we should use to determine them. The next two chapters pick up the issue prioritisation methodology.

Section 6 concludes.

2. Happiness, subjective well-being, and measurement

In this section, I distinguish between happiness and SWB, discuss how those concepts are usually measured, and explain why life satisfaction scores are a suitable (if not theoretically ideal) proxy measure of happiness.

I define happiness as a net balance of pleasant over unpleasant experience and, as such, understand it in a familiar Bethamite way.[13] I follow Crisp in holding that happiness is uni-dimensional, in that all sensations can be put on a single scale of pleasantness.[14] I assume there are no higher/lower pleasures.[15] Hence, the only two components to happiness are intensity (how pleasant/unpleasant something feels) and duration (how long the sensation lasts). Following Edgeworth, I think that a theoretically ideal measure of happiness would be the 'hedonimeter', which would measure the pleasantness of the subject's moment by moment experiences.[16] If we could plot a person's happiness at each moment, their total lifetime happiness would be the sum of all the individual moments.[17]

Philosophers have offered other definitions of happiness. The alternatives seem to be life satisfaction[18] (happiness consists in having a favourable attitude towards one's life as a whole), the emotional state view[19] (roughly, happiness consists in a propensity for being in a good mood), and 'pro-attitudes'(Feldman, 2010b) (roughly, happiness consists in a cognitive endorsement of various aspects of one's

---

[13] (Bentham, 1789)
[14] See (Crisp, 2006) For criticism of uni-dimensionality, see e.g. (Nussbaum, 2012).
[15] (Mill, 1861) is the original advocate the higher/lower pleasures distinction.
[16] (Edgeworth, 1881)
[17] This rules out, for instance, weighting the moments of your death more heavily than earlier moments of your life, or supposing that later moments of your life can retrospectively reduce how much happiness you experienced earlier.
[18] (Sumner, 1996) If happiness did consists in life satisfaction, the later claim—self-reported life satisfaction being a good proxy for happiness—would be almost trivially true.
[19] (Haybron, 2016)

life, rather than in net pleasant experiences). It is beyond the scope of the chapter to evaluate these alternatives. For the sake of argument, I assume the view I stated. That said, it is unclear if choosing an alternative definition of happiness would alter the practical conclusions discussed in section 5.

'Subjective well-being' (SWB) is a term used in social science—primarily, economics and psychology—as an umbrella phrase to refer 'ratings of thoughts and feelings about life' and comprises distinct components.(Dolan and White, 2007) An important distinction is typically made between *evaluative* measures of SWB, a cognitive reflection by the respondent on their life (or some part of it), and *hedonic* measures of *affect*, which capture the respondents' feelings or emotional states at a particular point (or points) in time.[20] Affect is sometimes split into *positive affect* and *negative affect*, which can each be measured separately and have different determinants.[21] The former refers to pleasant emotions, such as joy, contentment and elation, the latter to unpleasant emotions, such as sadness, fear and anxiety. A further distinction is sometimes made between, on the one hand, evaluative and affect measures, and, on the other, *eudaimonic* measures, which capture a sense of meaning and purpose in life, or psychological functioning.[22] In what follows, I will only discuss the first two more standard components—evaluation and affect. A number of different measures for each of these components have been proposed, some of which are mentioned momentarily; a fuller discussion is beyond the scope of this chapter.[23]

---

[20] (Stiglitz, Sen and Fitoussi, 2009)
[21] (Diener et al., 1999)
[22] (Dolan and White, 2007)
[23] See (OECD, 2013) Annex A for a list.

What is the relationship between SWB and happiness? Often, the terms 'SWB' and 'happiness' are used interchangeably. This is not only technically incorrect but also misleading. On the earlier understanding of happiness, happiness consists in affect.[24] The evaluative component captures a judgement, namely of how people feel about their lives, rather than an experience. Hence SWB comprises both the experiences of happiness and evaluations of life. I will use SWB when referring to both the affective and evaluative components; otherwise, I refer to specific components or measures by their names.

The 'gold standard' for measuring the affective component of SWB—i.e. happiness—which seems the closest currently available instrument to Edgeworth's hedonimeter, is the experience sampling method (ESM).[25] On this, participants are prompted to record how good/bad they are feeling at that particular moment one or more times a day and what they are doing. An alternative affective measure is the day reconstruction method (DRM), where participants break their previous day into episodes—a bit like scenes in a movie—and state their feelings and activities in each one.[26] The theoretical advantage of the ESM over the DRM is that it does not require participants to remember how they felt, which is significant in light of research showing how error-prone our memories are.[27]

---

[24] More specifically, the pleasantness part of affect—affect is often split into 'valence' (i.e. pleasantness), 'arousal' (i.e. excitement), and motivational intensity (the urge to move to/from a stimulus). (Harmon-Jones, Gable and Price, 2013)

[25] (Csikszentmihalyi and Larson, 1987) Potentially, at some future stage, happiness could be measured via brainwaves, which would be closer than ESM to Edgeworth's hopes for a hedonimeter.

[26] (Kahneman *et al.*, 2004)

[27] Kahneman et al. (1993) famously demonstrated the 'peak-end' rule, demonstrating we do not remember the average intensity of experiences, and memories are skewed by the most intense and final moments. See (Kahneman *et al.*, 1993).

Most of the literature on SWB has focused on life evaluations, more specifically on measures of overall life satisfaction.[28] Life satisfaction is usually found by asking, "How satisfied are you with your life nowadays?" on a scale from 0 "not at all" to 10 "completely". There are two explanations behind the greater comparative focus on life evaluation measures. The first is practical: it is simply much easier to collect data on life evaluations than on experiences. Participants usually answer questions about life satisfaction in less than 30 seconds and it can easily be included in existing population surveys.[29] By comparison, the ESM and DRM require more work from respondents—the former is intrusive and the latter takes respondents about 40 minutes to complete.[30] The second reason is moral. Life evaluations are sometimes thought by economists to be a measure of 'decision utility', what people choose to do. Economists have historically taken this to be of greater moral importance than 'experienced utility', how life is experienced, i.e. happiness.[31]

There is now a wealth of data on life satisfaction. It has recently become possible to say (a point that I return to in section 5) to what extent various outcomes cause an absolute increase in life satisfaction on a 0-10 scale, which is what we need in order to determine cost-effectiveness.[32] By contrast, to the best of my knowledge, there is insufficient research on affect measures to draw the same conclusions.

In light of this, my suggestion is to use life satisfaction scores as a 'proxy' for happiness. A proxy measure is one that is thought to track the item of interest that can be used when a more direct measure is unavailable. For a measure to be a good

---

[28] (Boarini *et al.*, 2012) p8.
[29] (ONS, 2011)
[30] (Kahneman *et al.*, 2004)
[31] The terms 'decision utility' and 'experienced utility' are from (Kahneman and Krueger, 2006)
[32] See (Clark *et al.*, 2018) If all we knew what that outcome X increase happiness, but not by how much it increases happiness, we cannot undertake cost-effectiveness.

proxy, it must have a close correlation with an item of interest. Use of proxies is standard and uncontroversial: for instance, it has been common to use GDP-per-capita or health metrics, such as Quality-Adjusted Life-Years (QALYs), as proxies for well-being. As noted earlier, Singer and MacAskill rely on such proxies.[33]

The question of how life satisfaction compares to income and health metrics as a proxy for happiness is postponed until section 5. Here, I set out some evidence that life satisfaction is a reasonable proxy for happiness.

Across countries, Diener et al. find a medium correlation (0.55-0.62) between life evaluation between affect balance on two different life evaluation measures.[34] At the individual level, Kahneman and Krueger find a moderate correlation (0.38) between life satisfaction and net affect.[35]

Boarini et al. find that affect measures have the same broad set of drivers as measures of life satisfaction—for instance, both are positively correlated with income, being employed, being in better health, being more educated, being married, having friends to count on, and feeling safe walking alone.[36] However, the relative importance of some factors changes. For instance, a given change in income has 40% of the impact on affect that it has on life satisfaction and feeling safe walking alone has twice as big an impact on affect as life satisfaction. These findings are broadly consistent with other analysis of the difference between the determinants of life evaluation and affect.[37]

---

[33] See footnotes 7 and 8.
[34] These are life satisfaction and the Cantril Ladder, see (Diener, Kahneman, *et al.*, 2010).
[35] (Kahneman and Krueger, 2006)
[36] (Boarini et al., 2012)
[37] E.g. (Diener, Kahneman, *et al.*, 2010; Kahneman and Deaton, 2010)

Deaton and Stone identify some cases where the measures are at odds, noting that the affect measures vary on the days of the week, improve with age, and respond to income only up to a threshold.[38] However, evaluative measures are correlated with income, even at high levels of income, and are often U-shaped in age, and do not vary over the days of the week.

What the above demonstrates is that affect and life satisfaction measures do generally go together. Hence, we can use the latter as a substitute for the former. However, some caution is still needed: the measures differ in the degree to which different things matter and sometimes whether those things are positive or negative. Therefore, if we use life satisfaction as our proxy for happiness when making cost-effectiveness assessments, we need to keep in mind whether, and to what extent, the proxy will foreseeably send us 'off course' and then correct for that accordingly.

3. Measuring happiness

There are long-standing doubts in economics about whether happiness can be or needs to be measured. According to Layard:

> In the eighteenth century Bentham and others proposed that the object of public policy should be to maximise the sum of happiness in society. So economics evolved as the study of utility or happiness, which was assumed to be in principle measurable and comparable across people [...] All these assumptions were challenged by Lionel Robbins in his famous book *On the Nature and Significance of Economic Science* published in 1932.[39]

[38] (Deaton and Stone, 2013)
[39] (Layard, 2003) at p2.

Robbins argued that there is no way to measure the magnitude of different individual's satisfaction with different outcomes. Even if introspection allows person A to say how they feel, "[i]ntrospection does not enable A to discover what is going on in B's mind, nor B to discover what is going on in A's."[40] Further, Robbins asserted that economics is the science of behaviour "imposed by the influence of scarcity" and in order to study that, it is only necessary to observe revealed behaviour and assumed individuals have a stable set of preferences, not to measure how people feel.[41] Hence, without the ability or need to measure happiness, happiness was out and preferences were in.

Angner suggests philosophers have absorbed these historical doubts from economics and the idea that happiness cannot be measured has since then become widespread in philosophy.[42]

However, as noted (section 1), despite these historical doubts, the last few decades have seen a cascade of research on SWB within the social sciences.

This and the next section both discuss important methodological concerns about the measurability of happiness and draw us into the philosophy of science. In this section, I explain both the *psychometric* approach to measurement that proponents of SWB measures rely on and why SWB measures are deemed satisfactory, according to that approach. The next section focuses on whether the SWB measures have interpersonal cardinality. My arguments in this section are largely unoriginal, as the measurement of SWB is now an extremely well-trodden terrain and I am able

---

[40] (Robbins, 1932) at p. 124
[41] (Robbins, 1932) at p. 16.
[42] (Angner, 2013b) cites (Crisp and Chappell, 1998) and (Fehige and Wessels, 1998) as representatives of this view.

to draw on an extensive literature. Nevertheless, this is normally an argument of two halves: philosophers of science tend to say why SWB can, in theory, be measured but do not explain the evidence supporting the conclusion that it is, in fact, being measured; social scientists tend to marshal the evidence without explaining why this shows happiness is successfully being measured. Although happiness is the item of primary interest, I will generally talk about SWB (happiness and life satisfaction) as the same concerns about measurement apply to both.

As Alexandrova (a philosopher of science) observes, social scientists, when pushed by philosophers of science on why they think SWB can be measured, will appeal to *construct validation*.[43] On the psychometric approach to measurement, you assume there are various *constructs*, which are particular attributes or phenomenon, such as intelligence, personality, or well-being. These constructs are latent, i.e. not directly observable. You also assume that there are *measures*, which are ways to elicit the observable indicators of the construct. To check if the measures are *valid*, that is successful in measuring the constructs, social scientists engage in a process of *construct validation*, where the measure of the construct is tested to ensure that it *behaves in the way we think it should*, given the researchers' existing understanding of the topic. Alexandrova notes that social scientists do not just declare their measures to be valid: it is obligatory to subject their measures to a battery of tests to validate them—more on this shortly.[44] The results need to be looked at in the round: it is too hasty to discard a measure if it has *some* counter-intuitive results but the measure, in general, behaves sensibly—perhaps there was a measurement error or, on further reflection, the counter-intuitive result is correct.

---

[43] (Alexandrova, 2017) at p. xliv.
[44] (Anna Alexandrova, 2016) p. 133.

This is perhaps analogous to philosophers accepting that true theories can have counterintuitive implications.

Angner suggests the suspicion among philosophers and economists about the measurability of happiness rests on the fact happiness is not directly observable.[45] However, he points out that latent constructs can be measured on the psychometric approach, so this does not provide an objection, in principle, to measuring happiness. Angner does not offer an argument for the psychometric approach, saying:

> It is of course in principle possible that these psychologists, economists, and medical personnel are all mistaken, and that things like attitudes, preference satisfaction, and blood pressure are impossible to measure even in principle, but I take this possibility to be too remote to be worth considering.[46]

As far as I can tell, philosophers of science have no *in principle* objection to the construct validation approach on which the psychometric theory of measurement relies. For instance, Alexandrova and Haybron claim is it a defensible approach to measurement: "[it] follows a coherentist spirit according to which measures are valid to the extent that they cohere with theoretical and empirical knowledge about the states being measured".[47]

Philosophers of science have raised different concerns. I'll briefly mention these worries and argue they are not particularly problematic for our purposes.

---

[45] (Angner, 2013b)
[46] Ibid. at p. 232.
[47] (Alexandrova and Haybron, 2016) at p. 1099.

Angner raises the objection that well-being might not consist in SWB, but in something more Aristotelian in character, and hence it is mistaken to say measures of SWB measure well-being.[48] This is not an issue here: we are trying to answer the question of whether purported measures of happiness do measure happiness, which is distinct from the question of what well-being consists in. While I am very sympathetic to hedonism (well-being consists in happiness) and welfarism (well-being is the only intrinsic good), I do not argue for those here, nor is doing so necessary for the point at hand. It's worth noting that a great many non-hedonists will, in practice, want to increase happiness; this is because either either they consider it is one of—although not the only—constituent of well-being or it is deemed instrumentally valuable for increase well-being.

Alexandrova and Haybron's concern is that attempts to measure well-being are theory avoidant; that is, they do not engage in enough philosophical theorising. They raise two specific concerns.[49] First, that social scientists will often attempt to argue that A rather than B is a better measure of *well-being* because A correlates better with various factors, e.g. income or good governance, than B does. They object—quite correctly—that the question of which is the right measure of well-being should be settled with reference to our best theory of what well-being is, rather than by appealing to the evidence.[50] This, similarly, is not a problem if we've already decided that we are focusing on happiness.

Their second concern is that social scientists let statistics, not theory, determine what the correct measures are in the first place. For instance, PANAS, a popular,

---

[48] (Angner, 2013b)
[49] (Alexandrova and Haybron, 2016)
[50] (Feldman, 2010a) makes a similar point, citing several cases where social scientists have mistakenly taken empirical evidence to settle conceptual matters.

20-item affect questionnaire, asks subjects whether they feel enthusiastic, interested, excited, proud, alert, attentive, etc. This list was reached by *factor analysis*, a statistical method which sorts out how many different clusters of items—'factors'—account for variance in the data and then shrinks the list to the most central. They note that certain states which are presumably important for well-being are left off this list, e.g. 'serenity', whereas unimportant ones, e.g. feeling 'alert', are left on. They say the use of factor analysis allows investigators to avoid the "hard theoretical questions about [...] which states are most relevant to well-being".[51] They don't state this, but the implication is that social scientists should be prepared to stipulate what goes into the questions on the basis of axiological concerns, rather than let a statistical method (e.g. factor analysis) mechanically determine this.

This concern *is* slightly problematic. It means that multi-item measures, which we might take as a measure of a given construct (i.e. happiness or life satisfaction) might not be as accurate as a more theoretically 'finely-tuned' multi-item measure. However, it is only *slightly* problematic. It is hard to believe we would get radically different practical implications with the sort of minor alterations they suggest.[52] What follows is that we can and should use SWB data, but be mindful of what might change if these measures were more theoretically finely tuned. We should not let the best be the enemy of the good.

Now we've accepted construct validation in *principle*, the next step is to marshal the evidence that advocates of SWB measures give for saying those measures are of high

---

[51] (Alexandrova and Haybron, 2016) at p. 1106.
[52] Also, as their concern is with factor analysis, this only applies to multi-item measures and not to single item measures. Hence, it wouldn't apply to life satisfaction or affect measures, such as the ESM, where individuals are just asked to report how good/bad they feel.

quality *in practice*. This topic has been extensively reviewed elsewhere and the aim here is just to convey the main points.[53]

In assessing the quality of a measure, the two characteristics that need to be evaluated are *validity*, the extent to which a measure captures what it is supposed to measure, and *reliability*, the extent to which measures give consistent results in identical circumstances (i.e. have a high signal-to-noise ratio). Reliability is necessary but not sufficient for validity. If I have a set of bathroom scales that produces a random number every time I step on them, they are not a reliable measure of *anything*. Suppose, next, I have a working set of bathroom scales but use them instead as a measure of *height*. They will be reliable—they give the same scores, assuming my weight does not change—but are not valid—they do not measure height. Reliability is more straightforward to test as this can be done statistically. Validity, on the other hand, is an evaluative judgement with no single test: we have to assess whether the measure behaves in the way we expect it to; this is why we have to look at the sweep of evidence.[54]

Before we turn to the relevant evidence, there are three more things to say about validity. First, it is evaluated for a particular measure. As we are interested in two constructs—affect and life evaluations—the aim is to show that measures of affect and life evaluation are valid measures of affect and life evaluation, respectively. Second, validity is not an all-or-nothing concept, but comes in degrees. Hence, different measures of the same construct may differ with respect to their validity: we might think the experience sampling method has higher validity, as a measure

---

[53] See e.g. (Diener, Lucas, *et al.*, 2010; Diener, Inglehart and Tay, 2013; OECD, 2013)
[54] (Diener, Lucas, *et al.*, 2010) at p. 75

of happiness, than the day reconstruction method. Third, assessing validity is an iterative process. To check validity, we start by testing the hypotheses that are most central to the measure. If the measure gives implausible results, we may declare the measure invalid. If it passes, we go on to test more peripheral matters. If the measure gives us unintuitive results here, it is no longer obvious whether we should revise our original theory or assume instead that our theory was correct and the measure invalid. Hence, we need to demonstrate that the measure gets intuitive, unsurprising results in central cases before we can rely on it to investigate other areas.

Reliability is usually tested in two ways: by internal consistency—whether the items within a multi-item scale correlate, or different scales of the same measure correlate—and by test-retest reliability, where the same question is given to the same respondent more than once at different times. To illustrate the concept, imagine we have a hundred imperfect mercury thermometers, we think some are broken and want to know which ones. One thing we would do is find out which thermometers are giving the same score. Another is to see if they keep giving the same score when it's equally hot (as opposed to, say, altering at random). Reliability is tested for statistically using correlations. Correlation is a matter of degree—it is between 0 and 1—and whether correlation is sufficient is a matter of whether it measures the standards of social science. 0.7 is generally considered to be the acceptable level; although we would expect lower levels of reliability for affect, given we expect mood to change quite frequently.

Regarding life evaluations, Bjornshov finds a correlation of 0.75 between life satisfaction and the Cantril Ladder (an alternate measure of life evaluation) in a

sample of more than 90 countries.[55] Test-retest results for a single item life evaluation measure tends to yield correlations of between 0.5 and 0.7 for a time period of 1 day to 2 weeks.[56] Michalos and Kahlke report that a single-item measure of life satisfaction had a correlation of 0.65 for a one-year period and 0.65 for a two-year period.[57]

Regarding affect measures, Diener et al. state that the positive, negative, and affective balance subscales of their Scale of Positive and Negative Experience (SPANE) have Cronbach's alphas (a measure of internal correlation) of above 0.8.[58] Krueger and Schkade report test-retest correlations of 0.5 and 0.7 for a range of different measures of affect over a 2-week period.[59]

Hence, the measures are deemed reliable enough—they are not just picking up 'noise'.[60]

We turn to validity. Although this cannot be assessed by simple statistical tests, that does not mean we cannot sensibly evaluate it. As mentioned, we have to assess the evidence in light of our best theoretical understanding of the construct at hand. There is a number of different types of validity.[61]

First, *face validity*—do respondents judge the questions as an appropriate way to measure the construct of interest? In the case of SWB measures, it is somewhat obvious this is the case, e.g. asking people whether they felt sad yesterday is a good

---

55 (Bjørnskov, 2010)
56 (Krueger and Schkade, 2008)
57 (Michalos and Maurine Kahlke, 2010)
58 (Diener, Wirtz, *et al.*, 2010)
59 (Krueger and Schkade, 2008)
60 For example, (OECD, 2013) discusses the evidence and reached the conclusion (at p46) the measures are sufficiently reliable.
61 (OECD, 2013) list three types at p47.

way to assess whether they felt sad yesterday. Participants aren't generally asked about face validity, but this can be tested by (a) response speed and (b) non-response rates: if people either take a long time to respond or don't answer, it suggests they don't understand the question.[62] Median response times for SWB questions are around 30 seconds for single item measures, suggesting the questions are not conceptually difficult.[63] Non-response rates for life satisfaction and affect were about the same measures of educational attainment, marital and labour force status, suggesting that people find those questions to be as comprehensive as the ones about SWB.[64]

Second, *convergent validity*—does the item correlate with other proxy measures for the same concept? Kahneman and Krueger list the following as correlates of both high life satisfaction and happiness: smiling frequency; smiling with the eyes (the "unfakeable smile"); rating of one's happiness made by friends; frequent verbal expressions of positive emotions; the happiness of close relatives; self-reported health.[65] An association between stress hormone cortisol and happiness has been found: those in bottom quintile of affect had 32% more cortisol than those in the top quintile.[66] None of this is supposed to be surprising—these are the sort of things we would expect SWB to correlate with.

Third, *discriminant validity*—whether measures that are supposed to be capturing different things are, in fact, measuring the same construct. Our background theory is that life evaluations and affect are distinct concepts. Hence, if our measures of

---

[62] (OECD, 2013) at p. 45-6.
[63] (ONS, 2011)
[64] (OECD, 2013)
[65] (Kahneman and Krueger, 2006)
[66] (Steptoe, Wardle and Marmot, 2005)

both gave the same results, we would conclude both measures were measuring the same construct (although we wouldn't know which one). This was already discussed in section 2 when we assessed whether life satisfaction and affect have different correlates and noted they did: to give one example, people report higher affect on the weekend but not higher life satisfaction. This is what we might expect: people enjoy their weekends more, but this doesn't change their evaluation of their life as a whole.

Fourth, *construct validity*—whether the measure performs in the world the way theory would predict. This is the main test of validity. The previous three types of validity can be seen as preliminary steps to indicate the measure could be expected to capture the phenomenon of interest: for instance, affect measures could have face validity but nevertheless have entirely implausible associations. A range of facts supports construct validity for SWB measures.

Kahneman and Krueger report that intimate relations, socialising, relaxing, eating and praying are all associated with higher levels of positive affect; conversely, commuting, working and childcare and housework are associated with low levels of net positive affect.[67] Higher incomes are associated with higher life satisfaction and affect—up to a certain point—all around the world, at both the individual and country level.[68] Health status, social contact, education and being in a stable relationship with a partner are all associated with higher levels of life satisfaction.[69]

---

[67] (Kahneman and Krueger, 2006)
[68] (Jebb *et al.*, 2018)
[69] (Dolan, Peasgood and White, 2008)

Life satisfaction successfully predicted suicidal ideation and suicide rates 20 years later in a Finnish survey.[70]

Looking around the world, the countries' life satisfaction is as we might expect. Stable, wealthy, well-governed countries come top: Finland, Denmark, and Norway lead the ranking with an average of around 7.5/10. War-torn poor states do badly: South Sudan, Central African Republic, and Afghanistan are the bottom three with an average of about 3/10.[71]

Major life events, such as unemployment, marriage, divorce and widowhood, are all shown to result in long-term, substantial changes to SWB, as shown in figure 4.1. These data are found by tracking the individuals in a cohort over time and controlling for other variables. The changes are not permanent—with the exception of unemployment, individuals eventually seem to adapt and return to their pre-event level of SWB.[72] The thought is that, as we would expect, people eventually get used to most life events, but unemployment continues to feel bad because it is associated with lower status.

[70] (Koivumaa-Honkanen et al., 2001)
[71] (Helliwell, Layard and Sachs, 2019)
[72] (Clark *et al.*, 2008)

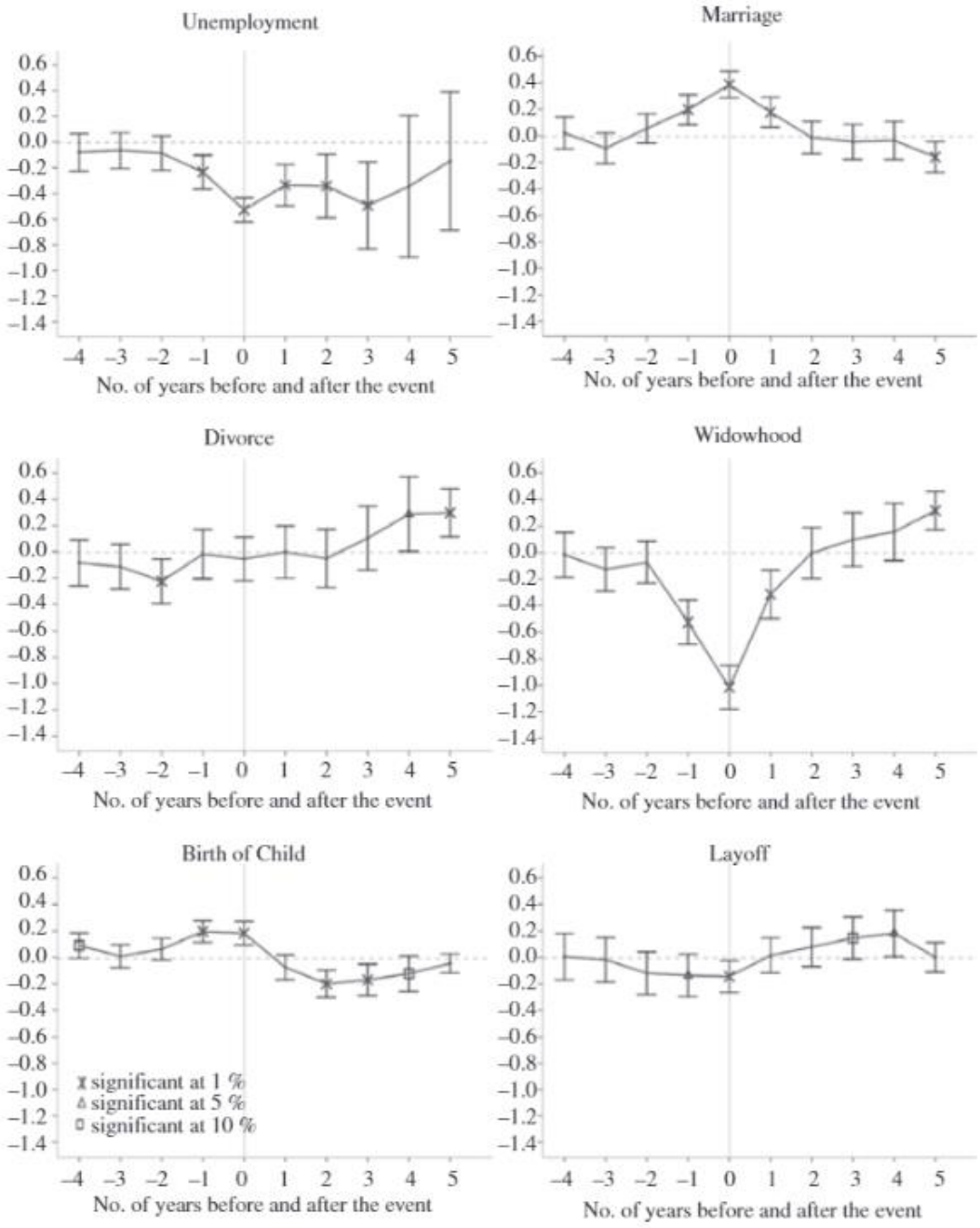

Figure 4.1. Lags and leads in life satisfaction in response to life events.[73]

Figure 4.2. shows that individuals only adapt somewhat to disability.[74] *Prima facie*, it is counter-intuitive for people to adapt even partially, and this is sometimes taken

[73] Reproduced from (Clark *et al.*, 2008) p. 235.
[74] (Clark et al., 2018)

as an argument against the validity of the measures. On reflection, partial adaptation is not so surprising: *becoming* disabled (e.g. losing the ability to walk) is a major shock but the state of *being* disabled, while worse than being non-disabled, is not as bad at becoming disabled. This is because one's aims, habits, and mindset adapt. One might doubt if individuals are adjusting how they use their scales while their actual SWB remains flat—this issue is picked up again in section 4.

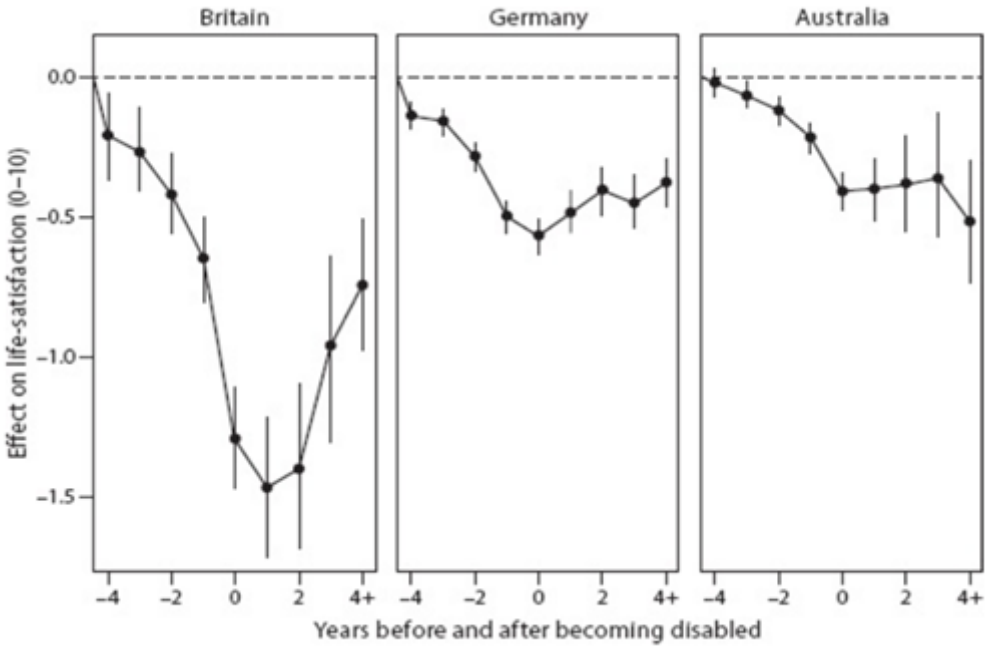

Figure 4.2. Adaptation to disability in different country datasets.[75]

One finding that is often used to suggest that the SWB measures are invalid is the so-called 'Easterlin Paradox'—the finding that while richer countries are more satisfied than poorer countries, and richer people in given countries are more satisfied, overall satisfaction seems to be broadly stable over time, at least in the developed world. This is displayed in figure 4.3.

[75] Reproduced from (Clark *et al.*, 2018) at p. 100.

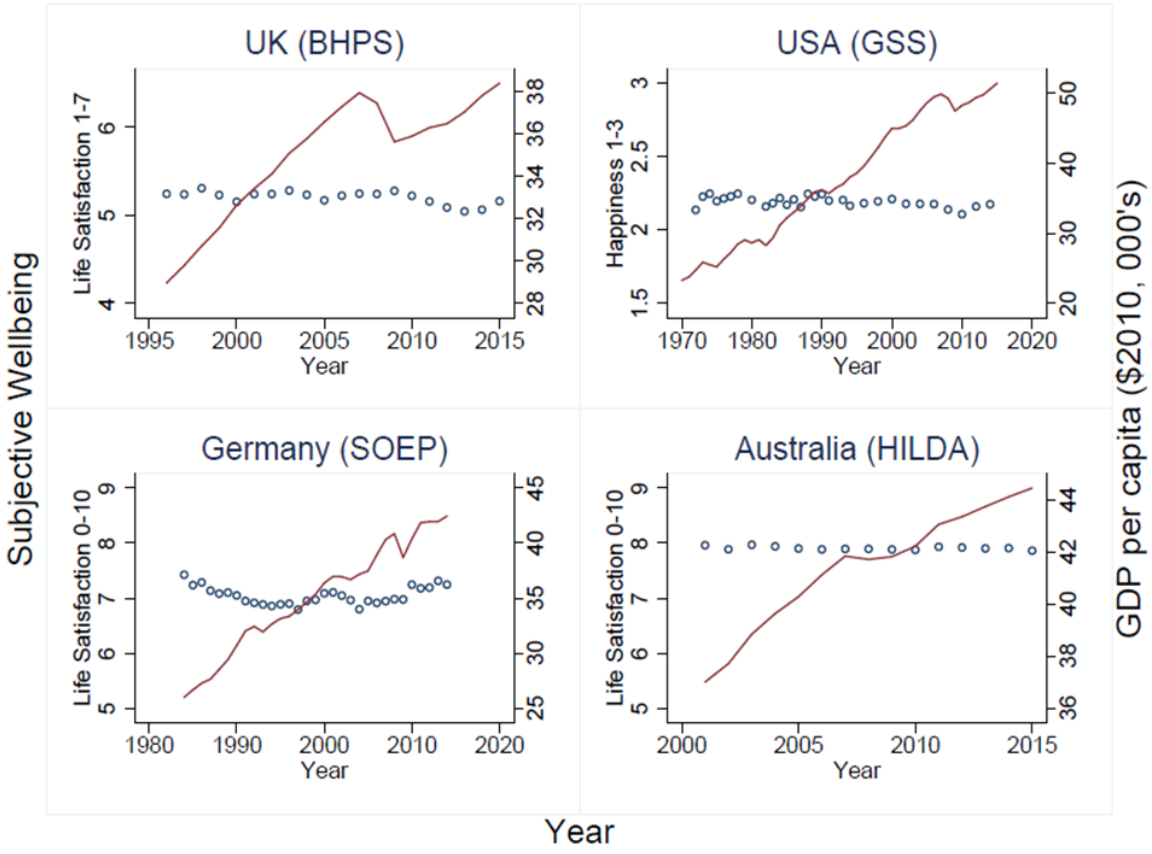

Figure 4.3. Change in subjective well-being and GDP/head over time.[76]

A critical response to SWB measures could be made as follows: "the Easterlin Paradox shows increasing overall economic prosperity doesn't increase SWB. But it's obvious increasing overall economic should raise SWB. Therefore, the SWB measures must be wrong".

However, such a response is too quick. First, the debate continues over whether the Easterlin Paradox holds: Stevenson and Wolfers argue it does not, Easterlin et al. reply.[77] Second, as Clark notes, a large body of research has found that individual SWB depends not just on the individual's own income, but also their income relative

[76] Figure from (Clark *et al.*, 2018)
[77] (Stevenson and Wolfers, 2008), (Easterlin, 2016)

to that of the reference group they compared their income to.[78] Thus, if I am wealthier than you, I should expect to have higher SWB. However, if my income rises, but the income of those I compare my income to also rises, these effects are cancelled out, leaving my SWB unchanged. Hence, when we apply some additional theoretical understanding—in this case accounting for social comparison—the results from the measures are no longer so counter-intuitive.[79]

Overall, on the construct validation approach, the case that the two types of SWB measures—life evaluations and affect—do measure what they set out in a very compelling way.

Before I moved on the comparability of SWB scores, it's worth noting that even if the measures are generally good, it does not follow they are flawless. Further, the fact there are some measurement issues is not sufficient to declare the scales invalid: if your bathroom scales gave you an implausible result, you would not thereby conclude from this that measuring weight is impossible. Regarding SWB, much has been made of studies which have found that seemingly irrelevant factors—such as finding a coin, being asked about your love life or your satisfaction with politics right before reporting your SWB—affect the results.[80] However, it's important to point out that the impact of such factors is relatively limited and has not been borne out by the wider literature. Diener et al. state that 60-80% of variability in life satisfaction is associated with long-term factors and the remainder with occasion-

---

[78] (Clark, 2016)

[79] Peter Singer raises the question of whether, if the Easterlin Paradox holds, it implies governments should not try to increase GDP. Increasing GDP can be indirectly useful: governments can raise more in taxes and use them to fund well-being increasing services. What would follow from the Easterlin Paradox is that governments should not expect that increasing GDP will, by itself, increase well-being.

[80] (Schwarz and Strack, 1999), (Deaton, 2012), (Tourangeau, Rasinski and Bradburn, 1991)

specific issues and measurement errors.[81] When large populations are surveyed, random errors will wash out. Such issues can also be minimised through careful survey design, although discussion of this is outside the scope of this chapter.[82]

4. Comparing happiness

This section focuses on the issue of whether the scale on which SWB is measured can be assumed to be *interpersonally cardinal*; that is, whether a 1-point increase in SWB for anyone will be equivalent to a 1-point increase for anyone else.[83] This topic is important because if our aim is to maximise SWB, we need to know how much better off we can make different people, which requires interpersonal cardinal comparisons of SWB. Despite the importance of this issue, as Kristoffersen observes, it rarely seems to be discussed explicitly: researchers tend to either treat the data as ordinal or cardinal and conduct different statistical tests as a result, without articulating their assumptions for doing so.[84] At present, there seems to be confusion about exactly what one needs to believe in order to accept self-reported SWB scores are interpersonally cardinal.[85,86]

To make progress, I do three things in this section. First, I set out the conditions which are jointly sufficient for the truth of:

---

[81] (Diener, Inglehart and Tay, 2013)
[82] See (OECD, 2013) and (Pavot, 2018)
[83] Equivalent in the sense of representing the same magnitude of change in the underlying psychological state being measured.
[84] (Kristoffersen, 2011) p103-4 provides a list of a dozen or so papers where the authors treat the scales as either cardinal or ordinal without specifying why.
[85] For instance, (Van De Deijl, 2017) conflates the discussion of two separate conditions—intrapersonal interpersonal cardinality and interpersonal intercultural cardinality—in his discussion.
[86] To be clear, here, I am not concerned about whether *well-being* has a cardinal structure. I will be assuming the components of subject well-being have cardinal structure and asking, given this, whether the self-reported subjective well-being scores are interpersonally cardinally comparable. For a discussion of well-being cardinality, see e.g. (Broome, 2004) ch5.

> The *Raw, Universally Cardinality* (RUC) thesis: 'raw' self-reported SWB scales are 'universally' interpersonally cardinal.

By 'raw' scores I mean the unadjusted numbers that people give—I'll come back to this in a moment. 'Universally' refers to the scales being comparable between people across different cultures and times. To explain this specification, note that one might think (say) that the scales are interpersonally cardinal in a given culture at a given time, but not across cultures or times. While we might not need the scales to exhibit universally cardinality if we only wanted to make, e.g. intracultural comparisons. As it is the strongest version of cardinality that we would ever need, it is useful to see if it obtains.

Second, having set out the conditions, I evaluate whether each condition is, in fact, met and thus if the RUC thesis is true. As noted previously, SWB itself is unobservable, which means we can't directly test for cardinality. Nevertheless, we can shed light on the plausibility of each condition through theoretical arguments and empirical tests. I confess I am unable to show, as it were, beyond all reasonable doubt, that RUC is true. For two conditions, the relevant evidence seems missing and the best we can (currently) do is some less-than-fully-satisfactory armchair theorising. I suggest what further work is needed.

Third, I consider what we should do if RUC is false. My proposal is that if one thinks the *raw* scores are not universally interpersonally cardinal, all one needs to do is perform the appropriate mathematical transformation such that the *transformed* scores will be universally interpersonally cardinal. Hence, worries about RUC are not, in principle, an objection against relying on SWB decision-making (assuming one would have used SWB data if RUC has been true).

Fourth, I suggest we assume RUC is true, at least until new evidence proves otherwise. Nothing on this assumption convenient, RUC seems *approximately* true (for a given, not specified, value of 'approximately'), and there doesn't seem to be a particular transformation we could apply that would take the raw scores closer to cardinality.

In what follows, I first set out different types of measurement scales, and then, in turn, make the points mentioned above. As was the case in the last section, while we are primarily interested in happiness, rather than life satisfaction, as almost exactly the same concerns arise for both we can be ecumenical and refer just to 'SWB'.

4.1. Units of measurement

Units of measurement are typically grouped according to their quantitative properties. The standard four-fold division is as follows.[87]

Nominal scales are used for labelling variables without quantitative information, for instance, gender or hair colour.

Ordinal scales contain variables which have a relative magnitude, such as the order that runners finish in a race—1st, 2nd, 3rd, etc.—but lack information about the relative difference between those magnitudes. Ordinal variables cannot be meaningfully added or subtracted from one another.[88]

Interval scales contain variables which have all the features of ordinal measurements but, further, the difference between measurements on the scale are

---

[87] (Edwards, 1964)
[88] In the sense that this does not tell you anything about the change in underlying quantity, e.g. how far apart the running finish.

equal-interval—this is the condition for *cardinality*. Celsius temperate is the classic example—the difference between each one degree of temperature is the same in terms of the change in thermal motion. What interval scales lack is a non-arbitrary zero-point on the scale—a location where there is no underlying quantity of what the scale measures. Celsius is an interval scale because 0 degrees Celsius does not mean there is no thermal motion. Ratios are not meaningful on interval scales: 10 degrees Celsius does not have half the thermal motion of 20 degrees Celsius.

Ratio scales are the same as interval scales with the difference being there is a non-arbitrary zero point. Examples of this include mass, time, distance and temperature when it is measured in Kelvins. Ratios are meaningful, e.g. 10 minutes is twice as long as 5 minutes.

Both interval and ratio scales have cardinality. The important question is whether SWB scales are cardinal, as opposed to merely ordinal, when they are used *inter*personally (across individuals). To see the problem, suppose the scales were interpersonally *ordinal*. In this case, if person A's SWB goes up by one on A's scale, and B's SWB goes down by one point on B's scale, we don't know whether the total SWB has gone up, down, or stayed the same. This is because ordinal scales do not represent the underlying difference in magnitude. Cardinality is necessary but not sufficient to get us to a scale of ratio quality. I do not argue for or assume a ratio scale.[89]

---

[89] A ratio rather than interval scale is required in at least two important cases. First, if we want to compare how much net well-being results from (a) saving lives vs (b) improving lives, we need to know how far the saved lives are about the point of zero well-being, which only a ratio scale has. Second, if we want to apply a non-utilitarian aggregation function when determining the value of a state of affairs. If one is e.g. a prioritarian and gives more weight to the worse off, to give lives extra weight you need to know how far they are from a zero point. Hence only having a scale of interval quality is a limitation. This limitation emerges starkly in chapter 7.3 when we compare saving to improving lives.

4.2. Conditions for the Raw, Universal Cardinality thesis

Depending on exactly what we want there to be cardinality of, more or less strict assumptions are required. For instance, one might think that, due to cultural differences, *intra*cultural interpersonal cardinal comparisons are possible, but *inter*cultural ones are not. Hence, we need a further assumption to get from the former to the latter. If the following six conditions are met, that is sufficient for individuals' scores to be universally interpersonally cardinal:

1. The underlying phenomenon of SWB (happiness or life satisfaction) has a cardinal structure
2. There is a linear relationship between self-reported and actual SWB
3. There is a consistent scale used over time for each individual ('consistent' in the sense that the same self-reported levels of SWB are used to represent the same actual levels of SWB)
4. Individuals have the same maximum and minimum capacities for SWB
5. ndividuals, in a given society, use the maximum and minimum points of their scales to refer to their maximum and minimum SWB
6. There is consistent scale use between societies

1—2 together entail scale use that is intrapersonally cardinal at a given time. Adding 3 entails scale use that is intrapersonally cardinal *over time*. Adding 4 makes scale use *inter*personally cardinal in a given society. Adding 5 makes scale use interpersonally cardinal across societies. Before I evaluate how plausible each condition is in turn, I will make two comments.

First, while the conditions are jointly *sufficient* for RUC, there are not all individually *necessary*—only some are necessary. It will be easier to explain this latter after more of the analysis has been completed.

Second, one argument that comes up often in the context of interpersonal cardinality is (what I'll call) the 'Washing Out' argument.[90] The standard version of this argument has two premises. First, that variations in individuals' scale usage and their capacities for SWB are random. Second, if these are random, so long as the surveyed populations are randomly constructed and large enough, any differences will statistically 'wash out' as noise and can be ignored. To illustrate this, if there are as many people with a (say) 10% greater capacity for SWB as those with a 10% smaller capacity, or as many people will self-report 1-point lower than average for a given level of actual SWB as will self-report 1-point higher, these differences will cancel each other out. The conclusion of the Washing Out argument is that we can treat the scales as interpersonally cardinal on *average*, even if we have doubts about doing this between any two given individuals.

While the argument is valid and, as we will see later, will greatly help us, the first premise is questionable and hence the argument may be unsound, as there *are* plausibly non-random differences. Hence, it is non-random differences that we must pay attention to as we examine the conditions, which we do now.

4.3. Assessing the conditions

> Condition 1: The underlying phenomenon of SWB (happiness or life satisfaction) has a cardinal structure

---

[90] For examples of this argument, see (Dolan and White, 2007; Bronsteen, Buccafusco and Masur, 2012).

If SWB lacks cardinal structure it would be confused to think that we could measure it on a cardinal scale.

Regarding happiness, it is introspectively obvious this has a cardinal structure. As Ng points out, we do think it is coherent to make claims of the following type, "being thrown in a bath of sulphuric acid would feel at least twice as bad as stubbing my toe".[91] Perhaps the acid is between 10 and 100 times worse. It is hard to say exactly, but this imprecision is not due to problems in the intrinsic nature of happiness or its measurement, but simply because it is difficult to recall a toe stubbing and compare it with any accuracy to an imagined acid bath.[92] It seems that we can compare the magnitudes of happiness in the two cases, which is only possible if happiness is cardinal. If happiness was ordinal then all that could be said was that the sulphuric acid was worse, but not worse by some amount.

It is unclear what evidence could be produced to show happiness is cardinal. Edgeworth regarded it as axiomatic—a 'first principle incapable of proof'—that no one could be infinitely sensitive and that the 'just perceivable increments' of pleasures would be the same for all persons and all pleasures.[93] If each just perceivable increment (hereafter 'JPI') were the same, then it would follow that happiness has a cardinal structure—one built out of JPIs. Note that we need not actually count how many JPIs have changed to get a cardinal scale: we might expect individuals can intuitively judge equivalent quantitative changes in sensations without knowing precisely how many JPIs they've gone up or down. This expectation seems reasonable in other cases. For instance, one instance of the

---

[91] (Ng, 1997)
[92] (Ng, 2008)
[93] Both quotations are from (Edgeworth, 1881) at p7

Weber-Fechner law in psychophysics is that the objectively measured sound pressure needs to roughly double to get the same increase in subjectively perceived loudness.[94] It seems unproblematic here to trust subjects to report when the same subjectively perceived increase in loudness occurs *without* requiring them to say exactly how many extra JPIs of loudness they have experienced.[95]

We can make similar claims about life satisfaction. As noted, life satisfaction is often taken to be a judgement. Arguably, there are JPIs for judgements too—people are presumably not infinitely sensitive in their evaluations. Alternatively, we might think that measures of life satisfaction are capturing the strength of a feeling of satisfaction; if there were the case, we could hold there are JPIs of this as well. For our purposes, given we're interested in life satisfaction only as a proxy for happiness, and both versions of what the underlying construct is are cardinal, it doesn't matter exactly what life satisfaction is.

Condition 2: there is a linear relationship between self-reported and actual SWB

It is often assumed that the relationship between self-reported and actual SWB is linear.[96] However, there is at least some reason to think that scale use might be systemically non-linear.[97] I outline and motivate the two most plausible non-

---

[94] (Jesteadt, Wier and Green, 1977) argue that the law doesn't quite hold, but this is unimportant for our purposes.

[95] It's not strictly necessary for all individuals' JPIs to have the same magnitude: so long as differences are randomly distributed, this will 'wash out'.

[96] Or, rather, is approximately linear. (Blanchflower and Oswald, 2004) point out that as individuals are only given a limited number of response categories and SWB varies (nearly) continuously, reported SWB will follow a step-function as individuals report the nearest response available. Hence, they assume the reports are roughly linear.

[97] If it is randomly non-linear, washing out will apply.

linearities, offer some general reasons to suppose linearity is more plausible, and then highlight specific issues with each non-linear option.

The first possibility is that there is a logarithmic relationship between reported and actual SWB. For definiteness, suppose that for every 1-point increase in *reported* SWB the level of *actual* SWB would double. This is shown in figure 4.4. In this case, the scales will not even be an *intra*personally cardinal measure; if this were the scale used, it would clearly be a mistake to take someone who reports 5/10 SWB and someone who has 9/10 and say their *average* SWB is 7/10.

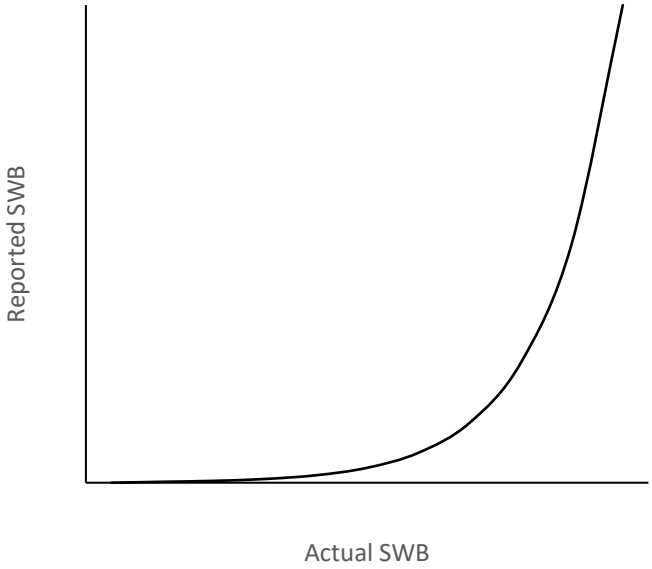

Figure 4.4. The logarithmic hypothesis

Why think that reporting of SWB works this way?[98] Some motivation comes from the Weber-Fechner law in psychophysics mentioned a moment ago. It's thought that the relationship between income and SWB works this way—income needs to double

[98] (Bond and Lang, 2018) in a recent (scathing) critique of happiness scales, seem to suggest this is a possibility; my information here is second-hand from a social scientist colleague as I found the paper mathematically impenetrable. Toby Ord (personal conversation) suggested that SWB scales were logarithmic and this was the reason to use income as the proxy for well-being instead.

to have the same increase SWB—so perhaps something similar occurs between actual and self-reported SWB.[99]

Ng suggests a different possible non-linearity, which is that the relationship between actual and measured SWB takes an arc-tangent form.[100] This means the distance in actual SWB increases at the extremes of the scale. Thus, the actual difference between a self-reported 9 and 10 (and 1 and 2) is greater than the difference between a 3 and 4, a 5 and 6, etc. This is represented in figure 4.5.

Ng's rationale is that SWB is theoretically infinite but measured on a bounded scale. Suppose someone becomes much happier than they thought was possible (as he has said occurred in his own case), they might be inclined to rate themselves as (say) 14 out of 10, based on what they thought the maximum of the scale *was*. This would prompt them to adjust their scale to accommodate all possible happiness values. However, if the individual kept a linear representation, given the potential range of values, Ng says this would problematically "compress normal changes in happiness values of say 20 per cent to a very small decimal value (e.g. 5.0010 vs. 5.0012)".[101] The apparent advantage of the arc-tangent is that it makes the scale's middle comprehensive while still allowing very high happiness scores to be represented at the top of the range. On Ng's model, those who say they are 10 out 10s are radically happier than the 9s. The same thinking applies to the bottom end of the scale too.

---

[99] (Kahneman and Deaton, 2010)
[100] (Ng, 2008)
[101] (Ng, 2008) at p. 257.

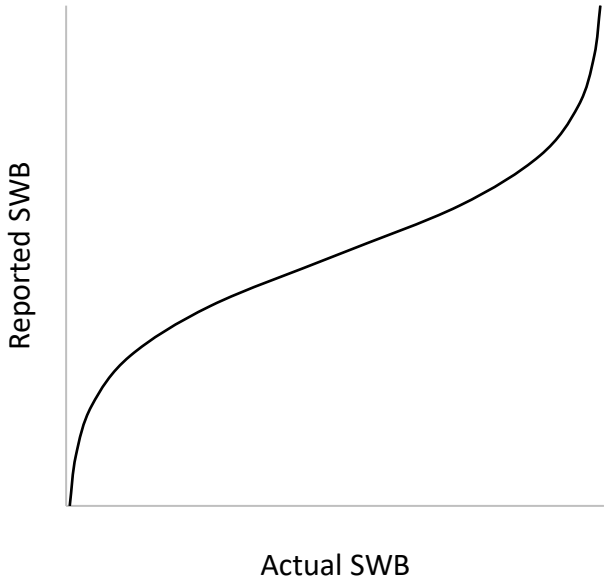

Figure 4.5. Ng's Arc-tangency hypothesis

The first argument in favour of linear reporting derives from Schwarz who argues that respondents try to work to understand what the researcher is asking them, as if they were talking to the researcher.[102] Applying this idea to SWB scales, Ferrer-i-Carbonell and Frijters suggest the most natural assumption for respondents to make is that they are being given a cardinal scale, as people are used to using cardinal scales, rather than ordinal or non-linear ones, in normal life anyway.[103] In support of this thesis, Van Praag argues that, in experiments, subjects do tend to interpret scales as being roughly cardinal: when subjects are asked to assign numerical values of between 1 and 1000 to five verbal labels (very bad, bad not bad, not good, good, very good), the consistent pattern across individuals is to space the words so they are numerically roughly equal-interval, i.e. individuals construct a cardinal scale.[104]

[102] (Schwarz, 1995)
[103] (Ferrer-i-Carbonell and Frijters, 2004)
[104] (van Praag, 1991)

Hence, it would arguably be perverse for participants to use anything but a cardinal scale unless they were expressly directed to do otherwise.

The second argument comes from statistical analysis. Krueger and Schkade conducted a test-retest of net affect one week apart.[105] We would assume that people's *actual* SWB would be expected to vary by about as much, week-on-week, whatever their level of actual SWB was.[106] If self-reported and actual SWB are linearly related, then this week-on-week change would be the same at different points on the scale. This is effectively what Krueger and Schkade find, suggesting the relationship is linear.[107] To make this finding consistent with the arc-tangency thesis, we would also need to suppose that those with very high or low levels of *actual* SWB would also have much higher variations in their SWB—this would make the variance in *reported* SWB appear to be linear up and down the scale.[108]

In a different test, Kristofferson found that the relationship between mental health scores and self-reported SWB data 'cannot deviate strongly from linearity'.[109] Kristofferson doesn't seem to make this explicit, but her argument relies on the assumption that mental health scores and *actual* SWB will have a linear relationship. If they did, the test would provide evidence of linear between mental health scores and self-reported SWB. This further assumption seems fairly plausible. Possibly, one could object that those who reach the ends of the mental health scale are relevantly different from those who are only *close* to the ends. But such people would have to be *very* different to support Ng's thesis, which is not what

---

[105] (Krueger and Schkade, 2008)
[106] Technically, we would expect 'homoskedastic errors'.
[107] (Krueger and Schkade, 2008) 18 note "assumption of homoskedastic measurement error could be violated, but the deviation is probably slight".
[108] A similar point can be made for the logarithmic thesis.
[109] (Kristoffersen, 2017) p845.

the evidence finds. It is not clear why one would think there should be a logarithmic relationship between mental health scores and actual SWB.

Turning to the plausibility of logarithmic self-reporting, one observation is that it would be unrealistically cognitively demanding to expect respondents to report in this way and, therefore, this is not what they do. If you ask me how happy I am on a 0-10 scale and I want to report this on a logarithmic scale, the first thing I intuitively do is to work out how happy I am on an arithmetic 0-10 scale. I then have to try to remember how logarithms work and convert my score on the arithmetic scale into one on the log scale.

Another issue with logarithmic reporting is it would imply that individuals think their SWB is implausibly low. Suppose, as many do, I say my SWB is around 7 out of 10. If I used the scale in figure 4.4, it implies my actual SWB (on a 0—10 arithmetic scale) is around 1.25.

Regarding arc-tangency, we need to distinguish two issues. One is whether individuals adjust their scales over time, such that SWB scales are not intertemporally *intra*personally cardinal—that is the question of condition 3 that we turn to in a moment. The other is whether arc-tangency or its reverse (where there is a smaller difference between the extremes than at the end) is more plausible from the armchair. Kristofferson notes that some people—around 10%—choose the highest scores in SWB surveys while others are reluctant to ever use the top score.[110] She cites Lau, who asked participants to recall an extremely good moment in their lives and asked those who did not give the highest score, a 10, why they had not done

[110] (Kristoffersen, 2011)

so.[111] 40% said that the top score is never attainable. What follows from this, Kristofferson argues, is that there may not be much of a difference in actual SWB between those who report a 9 or a 10—the differences in reporting are more of a reflection on the individuals' psychology. This runs exactly counter to Ng's thesis there would be radical differences between those who reported 9 vs 10. What follows?

Engaged as we are in armchair speculation, there do not seem to be stronger theoretical reasons to favour Ng's thesis or its reverse. Here we can appeal to the Washing Out argument: suppose we think some people self-report as Ng suggests and some people report the opposite way. We have no particular reason to assume one form of reporting is more prevalent than the other, hence the average reporting will still end up being *approximately* linear.

While we cannot prove linearity, it seems to be the only plausible assumption among the options.

> Condition 3: there is consistent scale use over time for each individual ('consistent' in the sense that the same self-reported levels of SWB represent the same actual levels of SWB)

Ng observes that happiness researchers seem not to have noticed the possibility that individuals can and do adjust their scales throughout their lives.[112] If individuals do this, self-reports are not *intra*personally intertemporally cardinal. I will first clarify what the problem is, make a few fairly speculative comments on whether it occurs and suggest how this could be tested.

---

[111] (Lau, 2007)
[112] (Ng, 2008)

Specifically, the concern is that some events happen to individuals which cause them to *rescale*, that is, they alter the levels of actual SWB that both the top and bottom of their self-reported SWB scale refer to. For instance, an individual, Sam, reports 7/10 SWB when playing tennis and has 7 actual SWB. Sam then becomes unable to walk and suddenly realises that his life can be much worse than he thought. Later, he is asked how he feels whilst playing his new hobby, chess, and he says he is 7/10.

Let's differentiate two possibilities. The first is that Sam's *actual* SWB is lower, say 6/10, but his *self-reported* SWB is the same. In this case, he has rescaled. The second is that his actual SWB is 7/10. In this case, Sam has not rescaled, but *adapted*—his SWB is back to where it was.[113]

If we want to accurately measure Sam's SWB then the occurrence of the second kind, adaptation, poses no issues for intrapersonal intertemporal cardinality. The first case, rescaling, is problematic. It means his self-reports represent *different* levels of actual SWB over time.

Suppose that Sam tells us he is 7/10. Should we assume he's rescaled, adapted, or done some combination of the two? As we can see from the earlier figures 4.1 and 4.2, people sometimes do report adaptation to live events, but this could, in theory, be entirely due to rescaling.

The first thing to say is that there are good evolutionary reasons to expect *hedonic adaptation* to occur, i.e. after a shock, happiness returns to where it was before.[114]

---

[113] A separate worry, which originates from (Sen, 1987) pp45-6 is whether the fact people do (or could) be happy in deprived circumstances. If we think such people are happy but have low well-being, that means well-being cannot consist in happiness. This is not our concern here.
[114] For more detailed accounts of how hedonic adaptation might function, see (Perez-Truglia, 2012), (Graham and Oswald, 2010).

The idea is that affective states are 'Mother Nature's' way to punish/reward animals for actions that increase/reduce our ability to survive and reproduce. Producing these sensations is costly in terms of energy, so to maximise effectiveness, hedonic adaptation is the 'rational' solution. Hedonic adaptation can occur at the cognitive level too—people change their views on things.[115] We wouldn't expect hedonic adaptation to occur in response to a situation that continues to be good/bad for the creature's survival; for instance, it would be disadvantageous to fully adapt to pain, as then pain would not be serving its warning function. As evidence of pain's usefulness, those with congenital immunity to pain, a rare medical condition, often end up severely damaging themselves.[116]

Second, if rescaling did occur, we would presumably expect to see it occur for all life events. As a matter of fact, this is not what we see. The life satisfaction scores in figures 4.1 and 4.2 above show that people report adaptation to some things—getting married and becoming bereaved—and not others—being unemployed and disabled. The simplest explanation is that when adaptation is self-reported, it has actually occurred. Someone who wanted to say individuals had rescaled but not adapted would need to supply a story about why divorce but not unemployment causes rescaling. I can think of no such story. In contrast, we can draw on our theory of hedonic adaptation to explain why it occurs in particular places and not in others. We would expect disability and unemployment to continue to be bad—the former continues to make life difficult, the second continues to feel shameful, and both are potentially isolating. We can see how it would be disadvantageous for people to be permanently sad after bereavement. Regarding relationships, it's worth noting that

---

[115] (Wilson and Gilbert, 2005)
[116] See e.g. (Udayashankar, Oudeacoumar and Nath, 2012)

while getting married (i.e. having a wedding) will cause a short-term increase in SWB (see figure 4.1), being in a relationship generally results in a permanent increase in SWB.[117] Getting married is merely having a large, expensive party, whereas spending time with someone you like continues to be good. So far, there is no need to suppose rescaling does occur.

Ng might respond, however, that I am simply not being inventive enough. Ng's suggestion is that individuals sometimes find themselves either more or less happy than they had previously reckoned *it was possible to be*; as a result, they expand their scales to accommodate the new possibilities. Perhaps disability and bereavement don't require rescaling because they are outside the bounds of what we *expected* when we formed the ends of the SWB scales. This still allows that events which are sufficiently good or bad that we *don't expect* could force rescaling. On the negative end of the scale, the intuitively worst thing is being tortured; on the positive end, perhaps achieving a moment of great victory—scoring a goal in extra time of the FA Cup Final, perhaps—or being high on some drug. People presumably have some idea of what these feel like, but it could be a surprise. It seems at least an open question whether these sorts of experiences cause rescaling.

Some further speculation. If the model is that only unexpected events cause rescaling, then SWB scales will underweight the best/worst experiences. It also implies that this won't lead to widespread loss of cardinality in the SWB data as such events presumably happen to only a few people. Further, it's unclear if these shocks would cause rescaling in typical reporting at all. Suppose some person, Sam,

---

[117] For timeseries on partnership, see (Clark *et al.*, 2018) on p. 80. For the long-term positive effect of divorce, we can see this as people returning to their pre-unhappy-marriage level of SWB.

becomes disabled in some grizzly accident, during which he feels more pain than he thought was possible. If asked about his SWB, he is tempted to say that, based on his previous usage of the 0-10 scale, he feels -5. As the scale can't go below 0, he expands the scope of his scale to accommodate his current sensation, i.e. what he would have called '-5' on the old scale becomes '0' on the new one. Some months later, he's then asked about his SWB. He thinks his actual SWB is just as high as it was pre-accident—what he would have said was 7/10. However, his scales have 'stretched' downwards. Now, we would he say he's 8/10 if he uses his *new* understanding of the scales. Yet, he realises that if he says he's 8/10, the person conducting the survey will think he's *more* satisfied than he was before. He's just as satisfied and wants to convey this, so he says he's 7/10, i.e. he uses his previous scale in order to make his old and new answers meaningfully comparable. Hence, if this is what happens, then scale use would be intrapersonally intertemporally cardinal anyway.

Particular empirical tests would shed light on this. One of these would be to ask people who have experienced major life events if (a) they felt better/worse than they thought possible and (b) whether they have changed from old reporting patterns and, if so, by how much and in which direction. With that information, it would then be possible to transform raw self-reported data to make them intertemporally intrapersonally cardinal. I am not aware that any such studies.

The other option which Ng suggests would be to develop an approach that measures SWB in terms of JPIs.[118] If each just perceivable increment has the same change in terms of SWB, then we could, in principle, count the exact number of JPIs that

[118] (Ng, 1997, 2008).

individuals are experiencing at different stages of their life, avoiding the issue of having a bounded scale altogether. As Ng notes, moving to a JPI approach would not only allow us to make intertemporal intrapersonal cardinal comparisons but also give us a *universally* interpersonal cardinal scale that would apply across times and cultures. There is not space to pursue how a JPI scale might work here but this is a potential fruitful avenue for further work.

This condition seems less plausibly met than the previous one was. It seems we can appeal to the evidence to rule out rescaling in some cases, but we cannot extinguish the worry that it does not happen in other cases. To settle this worry, further evidence seems necessary. I've given some theoretical reasons to suggest the rescaling may not occur or, if it does, it will not substantially change results at population level. My tentative conclusion is that we should assume this condition is met, at least broadly, until we find evidence that suggests otherwise.

> Condition 4: Individuals have the same maximum and minimum capacities for SWB

Assuming we've granted *intra*personally cardinality, this and the next condition are individually necessary and jointly sufficient to reach *inter*personal cardinality within a given society. To see this, note that there could be 'utility monsters' among us—those who can experience much more (or less) SWB than others.[119] If there were, then a one-point self-reported increase for A and B might mean 5 times the changes in actual SWB. Hence, the self-reported scale will not be interpersonally cardinal.

---

[119] (Nozick, 1974) at p. 41

The existence of utility monsters who can experience many times more SWB, however, seems to be extremely unlikely. Presumably, humans' capacities for SWB are determined by how our evolutionary history has shaped our biology and there's some level of sensitivity that has to be optimal for survival. To return to an example noted earlier, being immune to pain is an extremely problematic condition that would put someone at an evolutionary disadvantage.[120] What this suggests is that we should not expect huge differences.

However, we might expect there to be some differences and, indeed, there are: twin-studies indicate that about 33% of the variance in self-reported life satisfaction can be attributed to genetics.[121] At this point, I note there are two ways genetics could affect reports of SWB: (a) by altering the capacities for SWB, (b) by changing the levels of SWB. Suppose, as seems reasonable, individuals report their SWB by scoring it between the maximum and minimum experiences they consider it possible for them to have. Suppose A and B feel the same level of actual SWB, say level 7. A's genetics are such that his SWB goes from level 0 to 10, whereas B's goes from 0 to 8. On the self-reported 0-10 scale, B will report a higher number as B is closer to B's maximum. In this case, the scales are not interpersonally cardinal. The other possibility is that A and B have the same capacities, but A has different genetics and so has a higher *level* of SWB. In this case, the scales are still interpersonally cardinal.

Hence, whether genetics changes alter the capacities or levels of SWB is relevant to the question at hand. That said, I concede I have no argument to show genetic

---

[120] See footnote 116.
[121] (De Neve *et al.*, 2012)

variations alter levels rather than capacities, nor can I think of a test that would settle the matter beyond doubt.

For purposes of having a cardinal measure (and thus making decisions which rely on cardinality), genetic differences will only upset this if they occur non-randomly. If they occur randomly, they will 'Wash Out' across the surveyed populated. The problematic case would be if we thought, say, the Welsh had different genetics from the Scots, such as the two nations had different capacities, which meant their scores were not interpersonally cardinal.[122]

How do we proceed? My suggestion, once more from the armchair, is to suppose that we would not expect large, non-random differences in capacities of SWB on account of genetics. Self-reported scores will then still be roughly cardinal, even if not exactly.[123] I'll return to this point later, but if we do suspect there are non-random differences, we can, in principle, adjust the 'raw' reported scores so the transformed scores are cardinal: if we think the Welsh are 10% higher in SWB than the scales suggest, we can adjust the scales to account for that. It's not immediately clear what method we would use to ascertain there really was a difference; this is a topic for further work.

---

[122] Readers might wonder if biological differences between the genders cause differences in happiness. Reviewing gender difference in SWB, (Batz and Tay, 2018) found the evidence is 'highly inconsistent'. Where gender differences are found, the effect sizes are very different. The authors argue there are several explanations for the relationship gender and SWB (including biological ones) and disentangling them is difficult.

[123] Saliently, this response would not help us if we genetically engineered humans in order to give them higher SWB and then asked them for their self-reports. It might be possible to use JPI analysis to see how much higher their SWB is than ours, so we could correct for that in order to make cardinal comparisons between ordinary and enhanced humans.

Condition 5: individuals, in a given society, use the maximum and minimum points of their scales to refer to their maximum and minimum SWB

Keeping to the terminology, another problem is 'language monsters', those whose self-reported SWB score capture a different range and/or levels of actual SWB. An example is represented in figure 4.6.

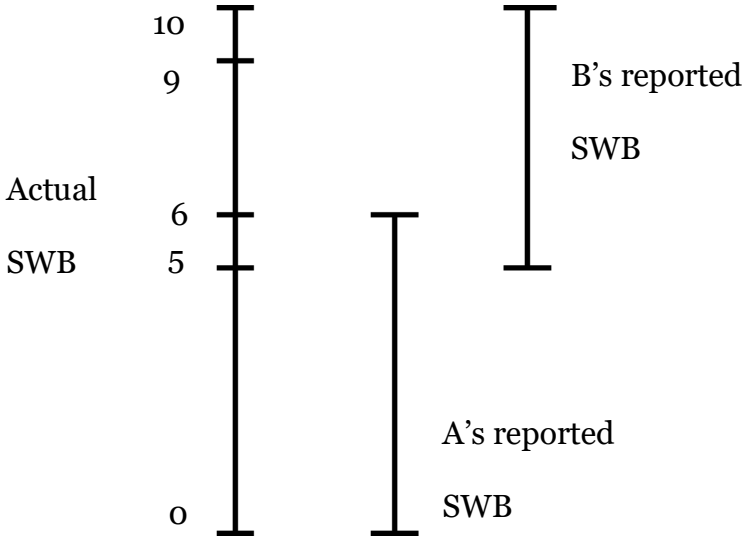

Figure 4.6. 'Language monsters'

As we can see, going from 0-10 on A's scale is worth 6 actual points of SWB, whereas only 5 for B. Hence a 1-point reported increase for B is worth 1.2 units of A's self-report scale, and hence the scales are not interpersonally cardinal.

A technical aside: specifically, the problem is if individuals self-reported scale do not have the same *range* of actual SWB. If A's scale had had a range of 5 units of actual SWB, then the two scales would be interpersonally cardinal: a one-point change for A would be equivalent, in actual SWB, as a one-point change for B.

Interestingly, it is thus not important for interpersonal cardinality that different people's scale represent the same *levels* of SWB.

This is why I stated the six conditions are jointly sufficient but did not claim they were all individually necessary. To see why they are not, note that we could imagine person C has 10 times the capacity for SWB than D does, but that C's self-reported scores only use 1/10th of the range of C's actually possibly SWB. Here, we would have interpersonal cardinality between C and D's self-reports but neither of conditions 4 or 5 would be met. If conditions 4 and 5 are met, however, people have the same actual range of SWB and use the full breadth of the range in their self-reports, which is sufficient to meet:

> Condition 4*: individuals self-reported scales use the same range of actual SWB (even if they use different levels of actual SWB), in a given society

Condition 4* that, along with conditions 1-3 and 6 are individually necessary and jointly sufficient for RUC. The easiest (only?) way to show condition 4* is met is by showing conditions 4 and 5 are met, which is why I arrange the conditions in this way.

Returning our consideration to condition 5, it seems unlikely that language monsters will be much of a problem *within* a given society, e.g. a nation-state. One reason for this is a reason given in relation to condition 3: Van Praag finds that when different people are asked to assign a numerical value to verbal labels (e.g. 'good') they give those labels similar values, suggesting that the words connote roughly the same emotional intensity to different people.[124]

---

[124] (van Praag, 1991)

Further, Van Praag offers the theoretical argument that the purpose of language is to transmit information between members of that language community. Thus, to be useful, words must have roughly the same meaning for any two individuals "[o]therwise, language would be no means of communication, and it is precisely that, which is the *raison d'être* of a language."[125] (emphasis in original). This is highly plausible, as evidenced by the fact that we regulate each other's language use: if I say, 'I stubbed my toe, this is worst I could possibly feel," I would expect the response, "Come on. That's not the *worst* you could feel."

Hence it would be surprising if individuals' scales differently substantially either in (a) the range of actual SWB they capture or (b) the maximum and minimum levels of actual SWB they refer to. If there is variation, presumably it would be random, which means the Washing Out argument applies again.[126]

Condition 6: there is consistent scale use between different cultures

There are mean-level differences in SWB across countries.[127] We might grant there will not be language monster issues within a given country (or linguistic community), but nevertheless, worry that cross-country(/cultural) comparisons are unintelligible on the basis of cultural differences.

---

[125] (van Praag, 1993) at p. 367

[126] Peter Singer raises the concern that extroverts, who report themselves as being happier than introverts, might not actually feel happier, but are simply using language differently (and this usage is arguably constitutive of their being extroverts). I am not sure if such a possibility could be ruled out by the empirical evidence, but it is possible to explain extroverts' higher self-reported happiness without appealing to language use. (Lischetzke and Eid, 2006) surveyed some possible explanations for extroverts' higher reported happiness; they then demonstrated through three studies that this difference can be accounted for by the fact that extroverts have better mood-maintenance abilities than introverts. I am unaware of any evidence indicating the particular possibility that Singer raises (this may be my ignorance). Without any such evidence, and given we have another explanation at hand, there doesn't seem to be a reason to suppose the worry Singer raises is true.

[127] (Helliwell, Layard and Sachs, 2017)

This has been an active topic of research and, as far as I can tell, it seems scale use should be treated as cardinal between cultures. An important initial point is that, as Diener et al. observe, mean-level SWB differences in a country could be a result of culture, but they could also be explained by differences in the situations between those countries.[128] Hence, we should not rush to assume it is culture that is doing the work. I note two recent intuitively compelling studies. First, Helliwell et al. examined a number of the predictors of life satisfaction across nations—e.g. income, levels of social capital, corruption—and found that "international differences in predicted values are entirely due to differences in their underlying circumstances" rather than different approaches to what constituents the good life.[129] Second, a study by Helliwell (and different colleagues) of immigrants moving from over 100 different countries to Canada found that, regardless of country of origin, the average levels and distributions of life satisfaction among the immigrants mimic those of Canadians.[130] This strongly suggests life satisfaction reports are primarily driven by life circumstances—if there were cultural differences in scale use there would not be such homogeneity of self-reports. Hence it seems likely that scale use is cardinal across nations.

Before moving on an assessment of the truth of RUC, I should quickly explain why the six conditions are jointly sufficient but not, as we might have suspected, all individually necessary. It could be the case that

---

[128] (Diener et al., 2017)
[129] (Helliwell *et al.*, 2009) at p 1.
[130] (Helliwell, Bonikowska and Shiplett, 2016)

4.4. Is the Raw, Universal Cardinality (RUC) Thesis true? What should we do if it isn't?

Let's sum up the discussion of the conditions. Conditions one, two, five, and six seem relatively unproblematic to assume (a cardinal structure for SWB itself; linearity between actual and reported SWB; common language use in each culture; intercultural comparability). For conditions three and four are not met beyond reasonable doubt (rescaling over time; individuals having the same capacities for SWB). However, and notably, we lack clear, compelling evidence they are not met, either. In short, some questions marks remain on those two. We cannot be very confident, on the basis of the present evidence, RUC is true.

What should we do if RUC is false?

One obvious line of thought runs as follows. Suppose we would have used SWB scores as our measure of happiness (or, I suppose, well-being) if RUC was true. We might think that, if RUC is false, we must abandon the use of SWB scores and return to something else to measure happiness—perhaps income or QALYs.

This is mistaken for the reason alluded to at the start of section 4: if we doubt the *raw* scores are universally interpersonally cardinal, what we should do is transform those raw scores such that the *transformed* scores are universally interpersonally cardinal. To elaborate on this, suppose someone was worried that condition 2—whether there is linear relationship between self-reported and actual SWB—was not met. All they would need to do would be to apply a local 'fix' related to that condition: whatever they think the deviation from linearity was, they could adjust the raw scores so the relationship between the *transformed* reported scores and *actual* SWB becomes linear. Hey presto, their worry has been addressed. We can't

'fix' the first condition: if SWB really lacks a cardinal structure, we must give up on having a cardinal measure. However, we can apply transformations to address any problems with the other conditions. Thus, if someone thought the Welsh were actually 10% happier than the data would suggest, they could make that change too. And so on. Hence, the person who wishes to use SWB scores should not give up on doing so whether or not RUC fails.

My concluding suggestion is that we should take RUC to be true, even though we cannot be very confident on this, at least until new evidence suggests otherwise. There are two reasons for this. First, to some level of approximation, RUC seems true. We might have doubts about the conditions, but none of them seems both (a) clearly not met and (b) not met in a way that would make the scores deviate substantially from universal interpersonal cardinality. For instance, while there may be some genetic differences that cause different maximum capacities for SWB, presumably there are not groups who have (say) 50% greater capacities. Second, and relatedly, while we could 'fix' the raw scores by applying a transformation, there doesn't seem to a strong justification for applying any particular transformation. As such, it doesn't seem we can take the raw scores any closer to universal interpersonal cardinality anyway.

5. Increasing happiness

The previous three sections argued that happiness can be measured through self-reports, individuals' happiness scores can be meaningfully compared, and that life satisfaction is a reasonable proxy for happiness. Three things follow from this.

First, the impact that different outcomes have on happiness can be determined (at least, in theory) by using self-reported subjective well-being scores; 'subjective well-

being' (SWB), recall, is an umbrella term which includes both happiness and life satisfaction.

Second, we can compare what looks best on SWB scores to the current suggestions made by Singer and MacAskill about how to increase happiness. Singer and MacAskill seem to have used health metrics (Quality Adjust Life-Years (QALYs) and Disability Adjusted Life-Year (DALYs) and income as their measures of happiness.[131] Let's call this alternative method the 'QALY+' approach, which I will return to shortly.

Third, if well-being consists in happiness, then we can get *close*—although, for reasons I will raise later, not all the way—to MacAskill's ambition to measure impact in terms of Well-being Adjusted Life Years (WALYs) instead of QALYs: we can use SWB scores.[132]

However, there are still two practical objections one could raise to using SWB scores instead of the QALY+ method to assess what increases happiness: first, there is not yet enough evidence on SWB to guide our decision-making; second, moving to SWB measures would not make a practical difference and is, therefore, unnecessary.

The only way to fully address those concerns is to show that we can crunch some numbers and that, when we do, it makes a difference. I postpone such analysis until chapter 7, where I provide a cost-effectiveness analysis, using SWB scores, showing a charitable priority that is different from the ones currently recommended by effective altruists. Before we get to that, I want to discuss, in the next two chapters, various methodological issues that point us towards the choice of priorities that I

---

[131] See footnote 7.
[132] See footnote 8

evaluate in chapter 7. My limited objective here is to provide reasoning that goes some of the way to countering the latter two objections.

In this section, I make two points. First, I note an advantage of using SWB scores over the QALY+ method to determine what increases happiness. Second, I state some of the latest SWB research and argue that mental health stands out as a potential priority, which is not evident if we rely on QALYs and income as outcome measures.

5.1 The problem of subjective weights

A general challenge, if one wants to assess cost-effectiveness, is to establish how much different outcomes contribute in terms of whatever common currency effectiveness is measured in. As WALYs aren't available, MacAskill suggests measuring the benefit of different interventions in terms of QALYs. For the moment, let's assume that the QALY is an accurate measure of happiness in the domain of health. QALYs are measured on a 0 to 1 scale with 0 equivalent to death and 1 to a year of full health. Different conditions are assigned different QALY weights: for instance, a year with AIDS without retroviral treatment is worth 0.5 QALYs—half as good as a year in full health.[133] If QALYs are accurate in terms of happiness, then having untreated AIDS would remove half of your net happiness for a year. QALYs would then be a good metric where we compared health outcomes.

The problem arises when we want to trade-off the impact of other outcomes, such as increased wealth, have on happiness. For definiteness, we could ask: how many

[133] (MacAskill, 2015) at p. 47.

years of doubled income are *equivalent,* in terms of happiness, to increasing someone' health from a QALY-weighting of 0 to 1 for 1 year?

If we measure happiness, an objective answer to this question can be derived: we can determine the impact that health and poverty each have on happiness. Unless we measure happiness, we're forced to make an educated guess about their relative impact. This is the problem of subjective weights: a factual question—how much do different outcomes increase happiness? —is being judged subjectively.

On what I've called the QALY+ method, we start with a health metric and then subjectively weight the relative importance of different outcomes relative to that health metrics. The obvious worry is that we'll provide flawed subjective weights.

5.2 SWB studies—some important results

Information about how much different things impact happiness can either be found (ideally) from randomised-controlled trials, or from large population surveys where statistical methods can be used to estimate the associations between various factors. Figure 4.7. is taken from Clark et al. It is the state-of-the-art and uses a national panel data set (i.e. the same people were surveyed each year), allowing individuals to be used as their own controls and the changes observed over time. [134]

[134] (Clark *et al.*, 2018) at p. 199.

	Effect on life-satisfaction (0–10)	*Total effect on the life-satisfaction (0–10) of others*
Income doubles	**+0.12**	**–0.13**
One extra year of education (direct effect)	**+0.03**	**–0.09**
Unemployed (vs. employed)	**–0.70**	**–2.00**
Quality of work (1 SD extra)	**+0.40**	—
Partnered (vs. single)	**+0.59**	**+0.68**
Separated (vs. partnered)	**–0.74**	—
Widowed (vs. partnered)	**–0.48**	—
Being a parent	+0.03	—
One physical illness	**–0.22**	—
Depression or anxiety	**–0.72**	—
Commit one crime	**–0.30 point-years**	**–1.00 point-year**

Figure 4.7. Life satisfaction effect different life-events on have self and others

If we use the life satisfaction scores, the most surprising lesson, relative to what we might expect from our folk psychology of SWB, is the comparative *un*importance of income and importance of mental health. From the table, we can see that not only does being diagnosed with depression or anxiety have about 6 times the effect on life satisfaction as a doubling of income, the aggregate effect of doubled income is roughly nil when the impact on others is accounted for; the right-hand-column figure indicates the effect that doubling one person's income has on others and suggests their loss is approximately the same size as the individual's gain. This result isn't particularly surprising: it's consistent with a wider literature (some of it mentioned in section 3) on social comparisons and how the effect of income on SWB is a substantially relative. I won't discuss the rest of the table as that is not necessary for our purposes. The above is developed world data, but it nevertheless indicates that mental health is a possible priority if we want to increase happiness.

Neither Singer nor MacAskill mentions mental health in their books on effective altruism.[135] Given the metrics they drew on, this result is not perhaps surprising. Not only does the above suggest mental health has a surprisingly large impact relative to income, different evidence that I now set out indicates mental health is relatively more important on SWB measures than it is on QALYs. Hence, QALYs turn out *not* to be an accurate measure of happiness in the domain of health.

Clark et al. provide the following explanation of how QALY weights are determined:

> In the QALY system, the impact of a given illness in reducing the quality of life is measured using the replies of patients to a questionnaire known as the EQ5D. Patients with each illness give a score of 1, 2, or 3 to each of five questions (on Mobility, Self-care, Usual Activities, Physical Pain, and Mental Pain). To get an overall aggregate score for each illness a weight has to be attached to each of the scores. For this purpose members of the public are shown 45 cards on each of which an illness is described in terms of the five EQ 5D dimensions. For each illness members of the public are then asked, "Suppose you had this illness for ten years. How many years of healthy life would you consider as of equivalent value to you?" The replies to this question provide 45N valuations, where there are N respondents. The evaluations can then be regressed on the different EQ5D dimensions. These "Time Trade-Off" valuations measure the proportional Quality of Life Lost (measured by equivalent changes in life expectancy) that results from each EQ5D dimension.[136]

---

[135] At the present time. Singer informs me that StrongMinds will be mentioned in the second edition of his book *The Life You Can Save* (Singer, 2019) due out in Autumn (and this is due, in some part, to my agitation on the subject).
[136] (Clark *et al.*, 2018) at p. 85.

Dolan and Metcalfe compare how individuals value the five dimensions of health in time trade-offs to the effect those dimensions have on SWB among people experiencing those states of ill-health.[137] This is displayed in figure 4.8. As we can see, the relative weights are quite different. To highlight a particular discrepancy, Dolan and Metcalfe report that subjects agreed hypothetically to give up as many years of their remaining life, about 15%, to be cured of 'some difficulty walking', as they would to be cured of 'moderate anxiety or depression'. However, from SWB measures , it is shown that 'moderate anxiety or depression' is associated with 10 times a greater loss to life satisfaction, and 18 times a greater loss to daily affect than 'some difficulty walking' is. On a moment's reflection, it is obvious anxiety and depression must be much worse for happiness than some difficulty walking, but this is not what we see in the QALY weights.

What explains this? The QALY method, as described above, is different from the SWB approach in two ways. First, the latter asks subject about their SWB, whereas the former, with the question "How many years of healthy life would you consider as of equivalent value to you?" leaves it open to respondents to answer in terms of whatever they value—what they value might not consist solely in (a component of) SWB.[138] Individuals presumably value SWB to some extent, but not all will only value that. The extent to which individuals do value goods besides SWB is an as-yet unresolved empirical matter and hence it is unclear how much of the disparity in results this difference in methodology accounts for.[139]

---

[137] (Dolan and Metcalfe, 2012)
[138] Arguably, the question leaves it open to answer not solely about prudential value (i.e. my own well-being) – I might say I want to live longer so I could do more good.
[139] (Adler, Dolan and Kavetsos, 2017) investigate hypothetical trade-offs between levels of SWB and levels of income, physical health, family, career success, and education. They found individuals prefer SWB to all the other attributes except health. This analysis doesn't tell us how much weight

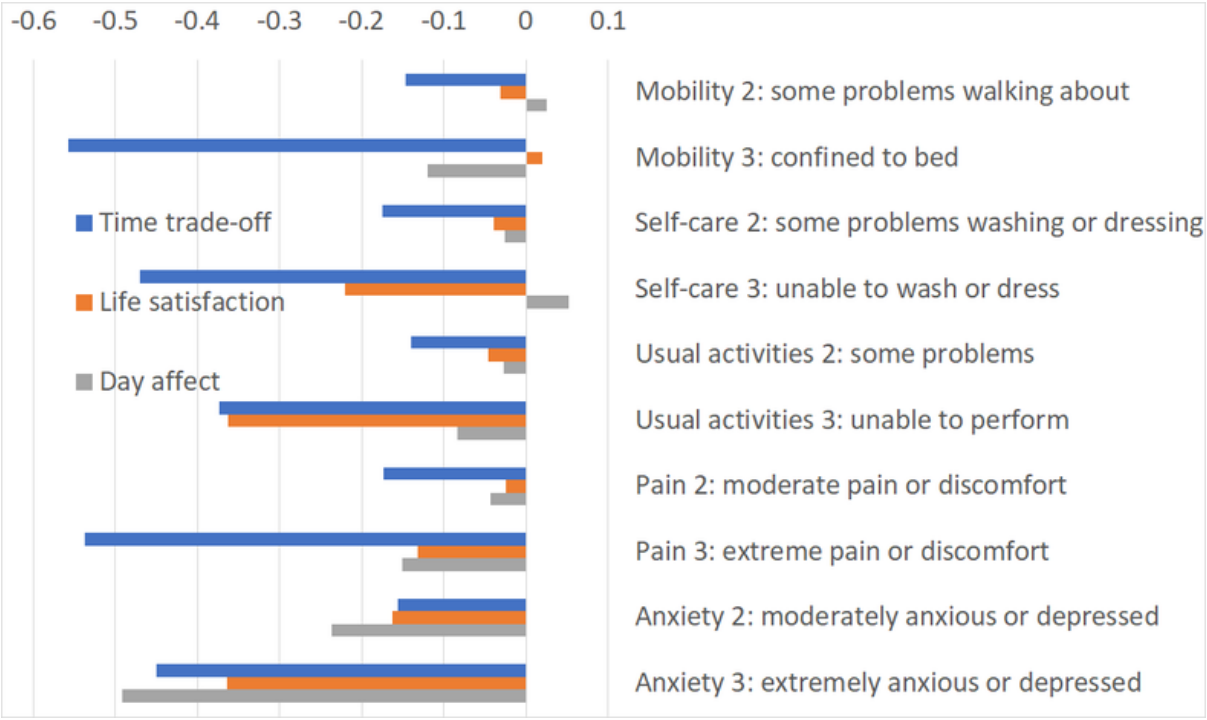

Figure 4.8. How life satisfaction and daily affect (0-1) are affected by the EQ5D, compared with weights used in QALYs.[140]

The second difference is that QALY weights are determined by asking people to *judge* how bad various health states would be, rather than by asking people who are experiencing those health states about their SWB and then *inferring* (via regressions) how bad they, in fact, are.[141] Psychological research into *affective forecasting*, how individuals expect to feel about future events, finds that individuals display an *impact bias*, overestimating the intensity and duration of future emotional states.[142] There are several reasons for this bias, such as *focusing illusions*, paying too much attention to easily imaginable details, and *immune*

individuals put on others goods vs SWB, or which of those goods are intrinsically valuable, only that they are, in practice, sometimes prepared to trade them off.

[140] Data from (Dolan and Metcalfe, 2012)

[141] This is the approach taken in (Clark *et al.*, 2018) at the wider SWB-literature in economics.

[142] (Wilson and Gilbert, 2005)

neglect, not accounting for the fact they will adapt to some conditions but not others.[143]

What seems to have happened is that, when you ask people to compare 'some difficult walking' to 'moderate anxiety or depression' they overweight the SWB impact of the former because it is easier to visualise—walking with a cane vs. feeling sad on the inside—and they haven't considered that they will adapt to the former but not the latter. The general problem with time trade-offs, however they are done, is that there is a difference between how important something seems when you are instructed to think about it compared to how it normally affects your experience.[144] As Daniel Kahneman pithily puts it, 'Nothing in life is as important as you think it is when you are thinking about it'.[145] This is why, if we want to know the SWB-impact something has, it is essential to ask people about their SWB in general, and then infer what impact their circumstances have on their SWB, rather than asking them what impact they *think* X or Y would have on their SWB.

Hence, if one used QALYs as a proxy for happiness, a key implication is that this would lead you to underweight the unhappiness caused by mental illness. I am not aware of equivalent research comparing SWB to DALY weights, but as DALY weights are also constructed by asking individuals to judge the badness of health states, the same concerns about affecting forecasting will presumably apply.[146]

---

[143] (Kahneman *et al.*, 2006), (Gilbert *et al.*, 1998)
[144] Hence, while QALYs are generally derived from members of the public making hypothetical judgments, they would not 'fixed' by getting those in poor health engage in time trade-off about their specific conditions; the concerns about affective forecasting remain.
[145] (Kahneman, 2011) at p. 400-1.
[146] (Gold, Stevenson and Fryback, 2002)

The preceding analysis should go some way to addressing two objections to using happiness data to guide (moral philosophers') happiness-related recommendations—that there isn't enough evidence to draw on and that, if we did, it wouldn't change our priorities. What it shows is there is some data on SWB and it shows us at least one new priority compare to the QALY+ approach, namely the potential importance of mental health.

Assuming we want to make people happier, given we now have a new methodology—SWB scores—to assess this and, with it, new evidence, this should prompt us to revaluate our priorities and see if we can find more effective ways to increase happiness. This, in turn, leads us to ask what method we should use, in general, for determining what our priorities are. Effective altruists have suggested such a method, a three-factor 'cause prioritisation' framework. In the next two chapters, we assess whether this framework is fit for purpose and, having suitably modified it, we take another look at what the happiness-increasing priorities are.

6. Conclusion

This chapter started with the realisation that while social scientists have been busy trying to measure happiness through self-reports, moral philosophers have not paid much attention to the social scientists' endeavours. I considered four possible objections to relying on self-reports to measure happiness. I argued the first objection can be met and made some progress towards address the remaining three. I am unable to say any more on the second objection in this thesis, but I aim to provide a compelling response to the third and fourth objections in chapter 7.

Chapter 5: How should we prioritise among the world's problems?

0. Abstract

In this chapter I set out what I call the 'EA method', the priority-setting methodology commonly used by members of the effective altruist (EA) community, identify some open questions about the method, and address those questions. I particularly focus on the three-factor 'cause prioritisation' framework part of the EA method. While the cause prioritisation framework is regularly appealed as a means of determining how to do the most good, it appears not to have been carefully scrutinised; its nature and justification are somewhat obscure. I argue that the EA method should be moderately reconceptualised and that, once it has been reconceptualised, we realise the method is not as useful we might have thought or hoped. Some practical suggestions are made in light of this.

1. Introduction

Let's begin with a long quote from William MacAskill, one of the effective altruism community's leaders, describing what I call the 'Effective Altruism method' or 'EA method', for short:

> Suppose we accept the ideas that we should be trying to do the most good we can with a given amount of resources and that we should be impartial among different causes. A crucial question is: how can we figure out which causes we should focus on? A commonly used heuristic framework in the effective altruism community is a three-factor cause-prioritization framework. On this framework, the overall importance of a cause or problem is regarded as a function of the following three factors

- *Scale*: the number of people affected and the degree to which they are affected.
- *Solvability*: the fraction of the problem solved by increasing the resources by a given amount.
- *Neglectedness*: the amount of resources already going toward solving the problem.

The benefits of this framework are that it allows us to at least begin to make comparisons across all sorts of different causes, not merely those where we have existing quantitative cost-effectiveness assessments. However, it's important to bear in mind that the framework is simply a heuristic: there may be outstanding opportunities to do good that are not in causes that would be highly prioritized according to this framework; and there are of course many ways of trying to do good within highly ranked causes that are not very effective. (Italics in original)[1]

As such, it seems the EA method has two main steps:

1. *Cause prioritisation*: comparing the marginal cost-effectiveness of different problems. This is done using the three-factor framework, i.e. assessing causes by their scale, neglectedness and solvability (these sometimes go by different names)[2]

[1] (MacAskill, 2018)
[2] 'Scale' has been called 'importance', 'neglectedness' called '(un/)crowdedness', and 'solvability' called 'tractability'. For different uses see, e.g. (MacAskill, 2015) chapter 10, (Cotton-Barratt, 2016), (Open Philanthropy Project and Karnofsky, 2014).

2. Of the causes which are most promising, move on to *intervention evaluation*: creating quantitative cost-effectiveness estimates of particular solution to given problems.

Although the three-factor cause prioritisation method is popular—central, even—to effective altruists' discussions about how to do the most good, there has been little analysis of the topic. There are a number of open questions about the method. I'll list and motivate a number of them.

First, what's the distinction between cause prioritisation and intervention evaluation? Following the quotation from MacAskill, it seems we can evaluate which causes are higher *priority*, that is, have greater cost-effectiveness, *before* making quantitative cost-effectiveness assessments of particular interventions. But it's not immediately clear how we can determine which causes are higher priority *prior* to considering, at least implicitly, the interventions we might use in each case and how cost-effective those interventions are.

Second, and assuming there is a distinction between the two steps, what reason is there for engaging in the cause prioritisation at all? We could just jump to intervention evaluation—e.g. assessing different poverty alleviation charities we could give to—and bypass cause prioritisation altogether.

Third, what role do the three factors—scale, neglectedness and solvability—play in helping us to determine a cause's priority?

Fourth, why are there three factors (rather than, say, four) and why are these particular factors used?

I aim to clarify matters by working through these questions.

The chapter is structured as follows. Section 2 suggests an initial distinction between cause prioritisation and intervention evaluation and explains how it is sometimes possible to determine causes are low-priority without examining particular. Section 3 considers how to evaluate plausibly high-priority causes, distinguishes two stages of cause prioritisation and explains why and how to use scale-neglectedness-solvability framework to determine the cost-effectiveness of causes at this second stage. Section 4 argues there is a very thin distinction between the second stage of cause prioritisation and intervention evaluation, suggests a slight reconceptualisation of the EA method, notes some worries that result from this, and asks whether this reconceptualisation is surprising. Section 5 makes some concluding remarks.

2. How and why to prioritise causes

It will help us to answer the questions posed if we realise that 'causes' and 'interventions' are just different words for 'problems' and 'solutions', respectively.[3] To be clear, if we want to solve a problem, we will eventually have to use some particular solution(s) to the problem. Hence, reframed this way, the query becomes whether and how it is possible to evaluate problems in some sense, 'as a whole', as a distinct process from evaluating *particular* solutions to those problems.

The explanation seems to be that we can evaluate problems *as a whole* if and when we can say something about the cost-effectiveness of *all* the solutions to a given problem. (From here, from stylistic reasons, I will usually talk of 'evaluating problems' instead of 'evaluating problems as a whole' even though the longer phrase

[3] After coming to realise this reframing myself, I subsequently discovered it had previously been used by (Dickens, 2016).

is more appropriate). In this case, we can evaluate the priority of a problem without having to look too closely at any of the *particular* solutions. Doing this is convenient because it allows us to quickly discard problems in which all the solutions seem cost-*in*effective and hence focuses our attention on investigating the more promising problems. There seem to be three situations where we can evaluate a problem and discard it as low-priority.

First, if the problem is not solvable by any means. For instance, we might say putting our resources (i.e. money or time) towards gun control in America is not a good altruistic decision. All solutions seem to involve political reform; opposition to gun control is sufficiently strong it's hard to imagine any political effort would succeed. Hence, it's unnecessary to get into the specifics of *any particular* solution/intervention to gun control because we are confident *none of them* will be as cost-effective as some alternative altruistic option we already have in mind (perhaps e.g. alleviating poverty). Another example would be if someone suggested building a perpetual motion machine and showed us their blueprints—we can rule out all particular solutions as being unfeasible without needing to look at the schematics for *this* version.

Second, if the problem will be solved by others whatever we do, our efforts will have no (or only a tiny) counterfactual impact. Some people think working to reduce climate change is a bit like this: there are now so many concerned individuals that my impact would be nearly nil, and I'll do more good by doing something else. A further, stylised example would be that a child has fallen into a shallow pond and someone else is already saving them—it won't help if you jump in too.

Third, if the problem is so tiny it's clear that putting any effort towards solving it would be a waste (compared to some already known potentially effective

alternative). Potentially, saving some rare species of beetle is like this. While it would (perhaps) be good if you did so, rather than doing nothing, you might conclude, without needing to look further at the details, the value will be so small that something else must be a higher priority.

Rather neatly, and perhaps unsurprisingly, the three types of situation each draw on one of the three factors used for cause prioritisation—solvability, neglectedness, and scale, respectively. As we're understanding intervention evaluation as the process of examining particular solutions to a problem, the foregoing analysis explains how and when we can engage in cause prioritisation without needing to get into intervention evaluation: sometimes we can 'screen out' entire problems quite quickly by noting something about *all* their solutions and concluding that none of them is comparatively cost-effective.[4] Given we're looking to do as much good as possible and our resources are finite, time spent investigating many different ways to solve unimportant problems is time wasted.

A potentially helpful analogy here is determining where to go on holiday. In the end, we have to visit a particular part of a country, rather than every part (or an average part) of a country.[5] Nevertheless, when thinking about our choice, we might start by seeing what we can say about countries as a whole (cause prioritisation) before considering the individual locations in more detail (intervention evaluation). If the country has features that make it unappealing—it's ruled by a dictator, they speak French there, it's very expensive, etc.—we can rule out going there without the need to look too closely at any particular places in those countries we could visit.

---

[4] If clarity is needed, I mean all the solutions that are actually possible, as opposed to, say, metaphysically possible.
[5] Except, I suppose, very small countries.

3. Prioritising causes using scale, neglectedness, and solvability

In the examples above, we could easily screen out those causes for one reason or another. Of course, we would also want our cause prioritisation methodology to tell us, of the problems which remain, which are the highest priority. Let's distinguish two stages of cause prioritisation. *Stage A* is where we screen out some of the problems, which we do by assessing all their solutions—this is the stage we have already discussed. At *stage B*, we determine the marginal cost-effectiveness of the problems which remain. In this section, I explain how the three factors (scale, neglectedness, solvability) combine to determine marginal cost-effectiveness of different causes and allow us to make progress at stage B.

MacAskill, in an explanatory footnote, offers more formal definitions of the terms used informally in the first quote and sets out how the three factors combine to determine the cost-effectiveness of a cause:

> Formally, we can define these as follows: Scale is good done per by percentage point of the problem solved; solvability is percentage points of a problem solved per percentage point increase in resources devoted to the problem; neglectedness is percentage point increase in resources devoted to the problem per extra hour or dollar invested in addressing the problem. When these three terms are multiplied together, we get the units we care about: good done per extra hour or dollar invested in addressing the problem.[6]

[6] I believe (Cotton-Barratt, 2016) was the one who first identified how the factors could be combined to assess cost-effectiveness. EA career's advice service (80000 Hours, no date) also uses this quantitative version of the framework.

We can write this as an equation:

> Scale x Solvability x Neglectedness = Good done/dollar
>
> (Good done/% solved) x (% solved/% increase in resources) x (% increase in resources/extra dollar) = Good done/dollar

Hence, scale, neglectedness and solvability are defined as factors which, when multiplied together, calculate the cost-effectiveness of whichever cause is being analysed.

It's worth briefly noting the existence of an earlier/alternative version of the three-factor framework that is qualitative in nature.[7] On the qualitative version, the same three factors were thought relevant to understanding a cause's cost-effectiveness—causes that were larger, more solvable and more neglected were presumed to be higher priority. However, there was no mechanism to trade-off the factors against each other or to combine them to determine cost-effectiveness, something this quantitative version allows. I use the newer quantitative version as it—unlike the older, qualitative one—allows us to determine the cost-effectiveness of causes and thus compare them to each other on that basis.

This brings us to the question of why there are three factors (rather than, say, four) and why these particular factors are used. This becomes clear once consider that causes generally have *diminishing marginal returns*: individuals will pick the 'low

---

[7] This is version is used by (MacAskill, 2015) in chapter 10 and (Wiblin, 2016). It seems to have been originally proposed by (Open Philanthropy Project and Karnofsky, 2014). In conversation, Hilary Greaves suggests that this version is arguably implicitly quantitative in nature, or at least used in that way: individuals applying the framework would assume something like the following: if cause A was X% larger in scale but Y% less neglected than cause B, the two causes would be equally cost-effective at a given ratio of X:Y. It might be true that framework has been in used this way, but given that no explanation has been offered as to why the factors could be traded-off like this, or what the appropriate ratios are, it's hard to see what justification there could be for using this framework to set priorities as opposed to just making educated guesses about the cost-effectiveness of particular interventions.

hanging-fruit' first, which means that initial resources put towards the problem will do more good per unit of resource than those added subsequently. This is represented in figure 5.1.

We can now point out the role each of the three factors plays. To determine the cost-effectiveness of additional resources, we need to know three things. First, the *scale* of the problem—how far up the Y-axis, which represents value, the line goes (how much fruit there is to pick in a given field).

Second, how much of the problem is solved for different amounts of resources, which is what *solvability* refers to (how easy it is to pick the fruit in this field and how this gets progressively harder). Combined with scale, this gives the shape of the cost-effectiveness line, which I'll call the *solvability line*. We cannot determine the solvability line with a single assessment, which is what we might have thought, but a series of assessments: given diminishing marginal returns, the fraction of the cause that is solved for a given amount of resources reduces as more resources are added. If there were instead constant marginal returns, that is, the cost-effectiveness was linear, then one assessment of the solvability line would suffice.

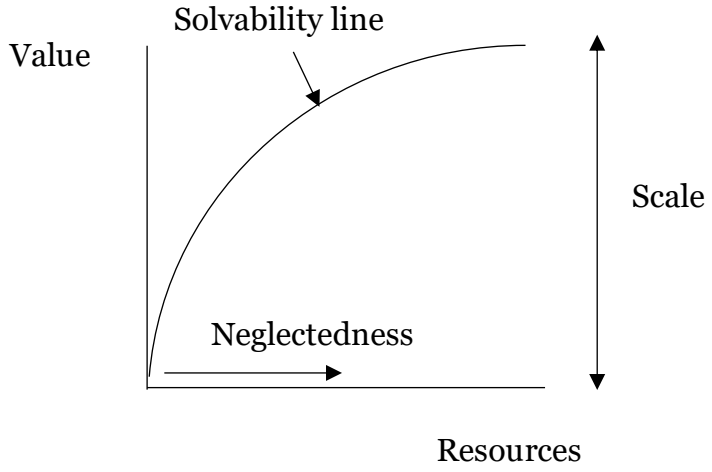

Figure 5.1.

Third, we need to know how many resources are being directed at the problem, which is what *neglectedness* captures. Imagine two identical problems, A and B, with the same solvability line. Suppose no one will try to solve A but many people will try to solve B. Assuming people will pick the low-hanging fruit first, then the cost-effectiveness of additional resources to B will be lower than A. The way we'd represent that on the graph is pushing along the X-axis the point at which we start counting marginal resources. This fact is illustrated in figure 5.2. overleaf.

Thus, we need to know the scale, solvability and neglectedness in order to correctly locate both 'where on the curve we are', so to speak—that is, where we should start counting the effectiveness of our marginal resources from—and where we would get to for the additional units of resources we contribute. If we know the three factors, we can determine the cause's priority (its cost-effectiveness). Note, however, that if there were constant marginal returns, i.e. cost-effectiveness was linear, we would only need two pieces of information to determine cost-effectiveness: first, the scale of the problem; second, a single assessment of solvability, i.e. the absolute number of resources required to solve X% of the problem.

Let's state the answers to the third and fourth questions posed earlier (what role do the three factors play? Why there are three factors (rather than, say, four) and why these particular factors are used?). To the third, the response is that the three factors are multiplied together to determine the cost-effectiveness of a cause; to the fourth, our goal is to determine cost-effectiveness, and these three are individually necessary and jointly sufficient for that task, hence the factors allow us to meet that goal.

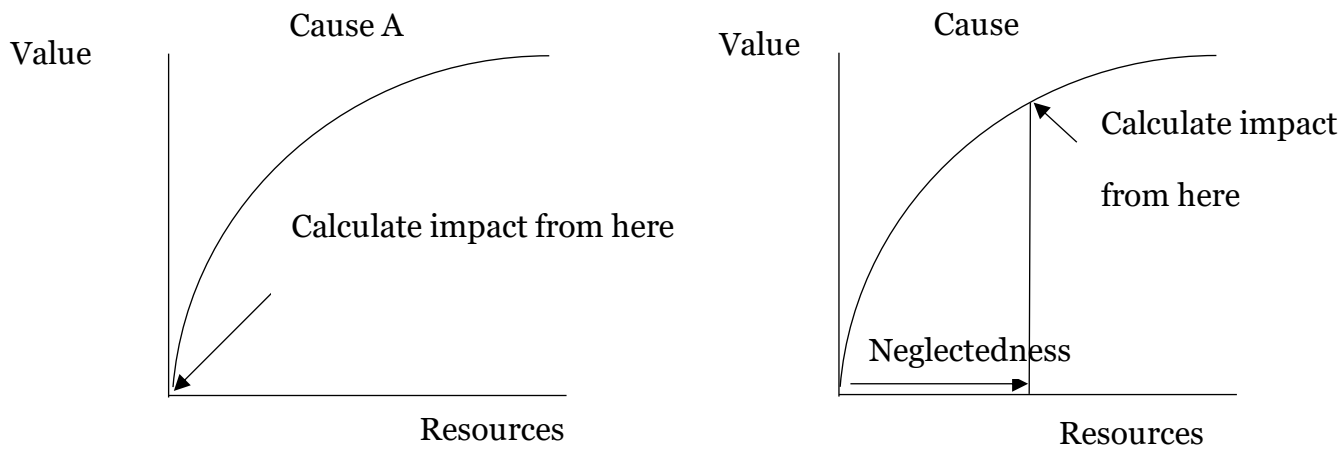

Figure 5.2.

4. Distinguishing problem evaluation from solution evaluation

We haven't yet got to the bottom of the first two questions (what's the distinction between cause prioritisation and intervention evaluation? What reason is there for engaging in the cause prioritisation in the first place?). In this section, I argue the distinction between stage B of cause prioritisation and intervention evaluation, if there is one, is very thin.

Suppose that we want to determine which of two causes, X and Y, is the priority. What we need to do at stage B of cause prioritisation is to plug in some numbers for each of scale, solvability and neglectedness and see what comes out the other end. But it's unclear there is any way to do this without considering, implicitly or explicitly, particular solutions to the problems at hand. Where else could the numbers come from? More specifically, as we're trying to do the most good, the relevant comparison is between the best solutions we are aware of for each problem, as opposed to (say) the median solution to the problem. The quality of our analysis will depend on our inputs to the formula being realistic, hence we want to have actual, particular solutions in mind, even if we are just intuitively weighing these up in our heads. Hence, what we're doing is evaluating particular solutions to given

problems. Yet, intervention evaluation is the process of evaluating particular solutions to given problems, so stage B of cause prioritisation isn't a distinct process.

We might think that although two processes are answering the same question—how cost-effective is a particular solution to a problem?—these are nevertheless somehow distinct. This is perhaps because we don't have to consider the whole problem (scale) when we evaluate solutions, or the resources going to the problem (neglectedness). But notice that when we evaluate interventions, if you know the cost to solve a given fraction of the problem and the value of solving that fraction, it's trivial to extrapolate those and work out the problem's scale; further, to determine the counter-factual impact of the solution, you do have to consider the resources that are going to the problem anyway.

Even though we may think we are referring to two different processes—stage B of cause prioritisation and intervention evaluation—both require the same inputs. What's more, not only do both produce an output in the same terms—good done per additional unit of resources—we'll presumably get the same answer whether we think we're doing one or the other: it would be very strange if our cost-effectiveness estimate of contributing extra resources to a problem is not identical to our estimate of how cost-effective the best solution(s) to that problem is. As a further indication these things are not distinct, notice that we could easily relabel any of the figures I introduced as representing the cost-effectiveness of causes, as representing the cost-effectiveness of interventions.

I imagine an advocate of the EA method might try to insist there is a meaningful distinction between stage B of cause prioritisation and intervention evaluation. There are two distinctions we might draw. First, the former is done intuitively—the three factors are combined in the head to make comparisons—whereas the latter

necessarily involves making explicit, quantitative cost-effectiveness assessments. Second, the former is an assessment of the best solution to the problem, whereas the latter is any assessment of some solution to a problem.

We can grant these distinctions, but it is still the case that stage B of cause prioritisation and intervention evaluation both consist in the same task: evaluating particular solutions to problems. Hence, if we assumed there was some deep difference between them, we are on thin ice. I note the second distinction is somewhat awkward: what follows from it is that if we use the three-factor framework to determine the cost-effectiveness of a problem and do this in our heads, it's 'stage B of cause prioritisation', but if we get out our calculators we're suddenly doing 'intervention evaluation'. Note that even if we do stage B in our heads, it is still a quantitative comparison of marginal cost-effectiveness that we are making.

What this analysis suggests is that we should conceptualise the EA method for setting priorities as having three steps:

1. 'Screen out' problems where it's clear all their solutions are cost-ineffective. This is done by appealing to one or more of scale, neglectedness, and solvability individually.
2. Make an intuitive cost-effectiveness evaluation of the most promising solution(s) to each problem. This is done by combining scale, neglectedness and solvability.
3. Make explicit, quantitative cost-effectiveness evaluations of particular solutions to problems.

These steps map onto (what I've called) stage A of cause prioritisation, stage B of cause prioritisation, and intervention evaluation, respectively.

We might have thought that evaluating causes consists only in assessing the problem as a whole; that is, all possible solutions to it. What follows from the preceding analysis is that once we've done the initial screening step and are trying to determine the relative priority of the causes that remain—i.e. how cost-effective resources will be in each—that analysis rests on an assessment of the cost-effectiveness of a particular solution(s) to the problem. Hence, while it may be more natural to make claims such as "poverty is a higher priority than climate change", it would be more accurate, though less elegant, to say instead "the best solutions to poverty we are aware of are more cost-effective ways of doing good than the best solutions to climate change we are of".[8] When phrased this latter way, it's clear that the cost-effectiveness of the solutions is doing the work in determining which of the problems is deemed the priority. Two concerns follow from this.

The first worry is that although the scale-neglectedness-solvability framework seems very sophisticated, it is just assessing the cost-effectiveness of particular solutions; our analysis here is only as good as the information we put in. Hence, if we have overlooked excellent solutions or wrongly estimated the cost-effectiveness of those solutions we (intuitively) considered, we will be mistaken about which problems are higher priority. If we thought that cause prioritisation via the three-factor framework resulted in a more holistic evaluation of the problem, we would not have this worry.

---

[8] Peter Singer asks whether a claim such as "poverty is a higher priority than opera houses" requires us to compare the best solutions in each case. I do not think we need to think carefully about the available solutions in each to defensibly make such a claim, but it still seems that we are (implicitly) appealing to the relative cost-effectiveness of the top solutions. To illustrate this, note how odd it would be to claim "poverty is a lower priority than opera houses", even if one thought the top poverty solution was more cost-effective than the top opera house solution, simply because there are some ways of doing good via opera houses than are more cost-effectiveness than some solutions to poverty.

The second concern compounds the first. The EA method invites us to divide the world up into cause 'buckets', look for the best item in each bucket, compare what we've found, throw away the buckets that seem to have nothing valuable in them, and then look further in the buckets that remain. The issue with this is that is if we have thrown away a bucket because we overlooked something valuable, the method discourages us from looking in that bucket again. The risk is that causes are determined to be low-priority prematurely: if we'd have looked for potential solutions longer, a good one would have been found. In the previous chapter, I argued mental health had been overlooked; this is perhaps part of the explanation.

The practical upshot of the analysis is as follows: if we want to find the most cost-effective ways to do good, and be confident we haven't overlooked the best options, we need to get stuck into the particular things we can do to make progress on each problem. Incorrectly believing that we can just take a quick look at a problem, consider its scale, neglectedness, and solvability, and thereby gain an accurate picture of a problem's cost-effectiveness, is liable to lead us to overlook things. It's these concerns which motivate, in the next chapter, proposing and testing a new approach to cause prioritisation that aims to overcome this issue.

To be clear, I'm not claiming these concerns are ones that (sophisticated) effective altruists are unaware of, or they demonstrate that effective altruist's prior prioritisation efforts are systematically mistaken and they must go back to the drawing board.[9] I am merely noting the concern and highlighting the limitations of the methodology.

---

[9] I say this despite the fact that, in the next chapter, I go back to the drawing board to (re)consider what the priorities are if we want to make people happier during their lives. This is, however, a combination of both (a) wanting to create and test a different prioritisation method and (b) because,

It's unclear if this proposed reconceptualisation is surprising or not. On the one hand, what I suggest seems to be in tension with comments others have made on this subject. For instance, Michael Dickens writes:

> The [three-factor] framework doesn't apply to interventions as well as it does to causes. In short, cause areas correspond to problems, and interventions correspond to solutions; [the three-factor framework] *assesses problems, not solutions*." (emphasis added)[10]

Hence, Dickens implies that it is possible to assess problems without assessing solutions *at all,* from which it would follow that cause prioritisation (stages A and B) is really distinct from intervention evaluation. As argued earlier, it's unclear how this could be possible.

Robert Wiblin says of the three-factor framework that:

> This qualitative framework is an alternative to straight cost-effectiveness calculations when they are impractical [...] In practice it leads to faster research progress than trying to come up with a persuasive cost-effectiveness estimate when raw information is scarce [...] These criteria are 'heuristics' that are designed to point in the direction of something being cost-effective.[11]

If the framework which we use for cause prioritisation is qualitative, then it must be distinct from the quantitative cost-effectiveness estimates we make of particular interventions. What seems to be going on here is that Wiblin is referring to the older, qualitative version of the three-factor framework I mentioned in the previous

---

as argued in chapter 4, I think effective altruists have been using a non-ideal measure of happiness and this alone prompts a reevaluation of the priorities.

[10] (Dickens, 2016)

[11] (Wiblin, 2016)

section. The newer version—the one we're using—is a quantitative framework where the factors do combine to produce 'straight cost-effectiveness calculations' and so the distinction between using the framework and making cost-effectiveness analyses of interventions disintegrates.

More generally, it is *somewhat* surprising that no one seems to have made it explicit that the priority-setting process should be broken into (at least) the three steps I have suggested. We might have expected someone to offer a clarificatory description of the EA method along the following lines: "first, we ignore the problems with no good solutions; then, we make intuitive judgments of how cost-effective the best solution to each of the remaining problems is; finally, we make some explicit, numerical estimates of those solutions to check our guesses".

On the other hand, I don't think anyone has *explicitly denied* priority-setting can work in the way I have just stated. Indeed, it seems clear, on reflection, that this is how we can and should approach the task. When we decide where to go on holiday, I presume most people make rough judgements about how much they want to visit entire countries, then consider particular in-country destinations in more detail and, at the final stage, compare prices etc. for the different options.

As a further comment against this being surprising, my analysis seems at least compatible with MacAskill's. In the opening quote, MacAskill states the three-factor framework allows us to make comparisons between causes even if we lack numerical cost-effectiveness assessments. What MacAskill is perhaps claiming is that the scale-neglectedness-solvability framework allows us to do step 2—make an intuitive comparison between problems—without having to go as far as step 3—making explicit quantitative cost-effectiveness assessments; his use of the word 'heuristic' seems to support this interpretation. As it happens, it seems plausible that we can

and do use three-factor framework in our heads to construct cost-effectiveness lines for different problems: we think about the size and shape of the curve and adjust where we are on the curve to account for what others will do.

As such, it's not obvious whether I am suggesting something different from what others thought of making explicit something that was previously implicit.

5. Conclusion

I began this chapter by asking: 'what's the distinction between cause prioritisation and intervention evaluation?' This was motivated by the apparently odd suggestion that we can do the former before the latter. I've argued there is a sense in which we can evaluate causes(/problems) as a whole: if and when we can evaluate all their solutions – this was stage A of cause prioritisation. If intervention evaluation requires the assessing of particular solutions in some depth, then stage A of cause prioritisation *can* happen prior to intervention evaluation. Both stage B of cause prioritisation and intervention evaluation require the cost-effectiveness of solutions to a given problem to be determined. If we stipulate that stage B of cause prioritisation can be done intuitively, whereas intervention evaluation requires the writing down of some numbers and then crunching those, the former can be also prior to the latter; I noted this distinction is pretty flimsy. If we don't stipulate this 'in-the-head' vs 'on-paper' distinction, then the two are the same process and hence stage B of cause prioritisation is not prior to intervention evaluation.

The second question related to why we should engage in cause prioritisation rather than leap straight to intervention evaluation. Following what we just said in the last paragraph, if we can evaluate causes as a whole—i.e. what happens at stage A—that

can usefully save time. We can now see there's not much difference between starting at stage B of cause prioritisation or with intervention evaluation in any case.

In this chapter, I've set out to address some of the outstanding questions about how the EA priority-setting method functions. What I've suggested here is a modest reconceptualisation of the EA method. I have not tried to argue the EA method is mistaken in some deep way, because it is not. Rather, I have tried to clarify something that seemed plausible but the details of which were murky. A practical conclusion emerged from the reconceptualisation: while we might have thought the three-factor cause prioritisation method assessed problems 'as a whole', this is only partially true: once we've sifted out unpromising problems (ones with no good solutions), our analysis of how problem A compares to problem B is nothing more and nothing less than a comparison of the best solutions to problems A and B that we've considered. Hence, if we want to find the most cost-effective ways to do good, we need to look carefully at these solutions.

Chapter 6: Finding ways to make people happier—developing and deploying the cause mapping method

0. Abstract

This chapter develops the three-factor cause prioritisation framework approach popularised by effective altruists. On my proposed method, 'cause mapping', the prioritisation process is broken down into several steps which are then systematically worked through to organise and identify potentially high-impact altruistic actions. I motivate and explain the method in general. I then apply it to the question of how philanthropists can most cost-effectively improve lives, that is, making people happier during their lives. The result is a long-list of potentially high-impact options; further empirical investigation will be required to determine which of these is the most cost-effective.

1. Introducing cause mapping

In the preceding chapter, I examined the priority-setting method that is typically used by effective altruists (EA) to determine which of the world's problem we should focus on if we want to do the most good; I dubbed this 'EA method'. The EA method seems to have two steps. The first is to evaluate causes (i.e. problems), for instance, poverty, mental health, factory farming, using the 'three-factor cause prioritisation framework', which involves assessing the scale, neglectedness and solvability of those problems (the details of this assessment are not important here).[1] This first step, it seems, is supposed to be prior to, and relevantly distinct from, the second step: making cost-effectiveness estimates of particular interventions (i.e. solutions) to given causes. I

[1] See (MacAskill, 2015) at chapter 10, (MacAskill, 2018).

argued that this conceptualisation of the priority-setting process isn't quite right and the EA method is better conceived of as having these three steps:

1. 'Screen out' problems where it's clear all their solutions are not cost-effective. This is done by appealing to one or more of scale, neglectedness, and solvability individually.
2. Make an intuitive cost-effectiveness evaluation of the most promising solution(s) to each problem. This is done by combining scale, neglectedness and solvability.
3. Make explicit, quantitative cost-effectiveness evaluations of particular solutions to problems.

If we've done steps 2 or 3 for any two problems, we can then say which of those two problems is higher priority (in the sense of having the more cost-effective solution). While we might say 'problem A is a higher priority than problem B', e.g. poverty is a higher priority than climate change, what is ultimately being compared is the cost-effectiveness of particular solution to each problem, namely the solutions that seem most cost-effective in each case.

Given this, the natural next question to ask is how we can do a good job of finding the most promisingly cost-effective solutions to problems. I am unaware of anything written within philosophy or among effective altruists on this topic.[2] A conspicuously bad approach would just be to pick, at random, actions we could take to solve a given problem and estimate their cost-effectiveness. I presume we want a method that will

[2] Perhaps this issue has been effectively resolved in some other area of academia. If so, I am nevertheless unaware of it.

allow us to organise our thinking, impose some structure on the task, and help us work out what the different things we can do are.

In this chapter, I develop a novel (but quite simple) approach, *cause mapping*, which does this. I explain and motivate it in general terms in this section. In the next sections, I put this method to work where the aim is improving lives, increasing well-being whilst people are alive, where well-being is taken to consist in happiness.

I'll now explain how cause mapping works; this requires the specification of some terms to differentiate different parts of the process. First, we list the *primary causes*, the problems we want to ultimately solve. Second, we list the *mechanisms*, the different types of methods that make progress on the primary causes. Third, we list the *obstacles*, the barriers stopping those mechanisms from being used. Each combination of a mechanism with an obstacle gives us a *solution*, a particular action we can take to do good. Hence combining the different mechanism-obstacle pairs gives us a list of solutions. By looking at the solutions and seeing what shared obstacles they have, we can then form a list of *secondary causes*. By seeing what obstacles there are to the secondary causes, we can also list some *meta-causes*.

This process may seem abstract, so an example of one part of the map will help. Mental health would be a primary cause. Providing psychotherapy for mental health is a mechanism for that primary cause, i.e. it improves mental health if it is used. However, having established a mechanism for the primary cause, we can then ask what obstacle(s) is preventing that mechanism from reaching everyone who would benefit from it. Here, money is the obvious obstacle. Hence one solution (for a philanthropist) is to fund psychotherapy for mental health; the solution combines a mechanism with a way of allowing that mechanism to be used. There are a number of specific types of psychotherapy that could be provided, so we can group these together and say that

generally ‘increasing access to psychotherapy’ is a secondary cause. To note the distinction between primary and secondary causes, the former refers to the type of problem we want to solve, the latter to an action (or type of action) we take to solve it. Regarding meta-causes, we might think encouraging others to give more to charity in general is a better way of increasing access to psychotherapy than funding it ourselves. Hence ‘encouraging altruistic behaviours’ would be a meta-cause—it is causally ‘upstream’ of actions which ultimately do good and has an impact indirectly through changing the behaviour of other agents.

Attempting to list all the possibilities at each step—all the primary causes, all the mechanisms, etc.—is clearly unrealistic, hence that is not what I suggest. Instead, I propose only to list the priority primary causes and their main mechanisms and obstacles. How are the priority primary causes determined? Through the same method used in the first step of the (reconceptualised) EA method above: we rule out unpromising primary causes using our intuitive judgements. Specifically, we see if any of the solutions to the primary causes seem more cost-effective than the most cost-effective solutions we are already aware of. The aim is to do the most good and our best existing option(s) ‘set the bar’: the aim is to clear it with something even better. Hence, there’s no point listing the mechanisms and obstacles that apply to unpromising primary causes. A similar process is applied thereafter: I won’t list all the mechanisms and obstacles to the priority primary causes, only those that seem they could lead to the most cost-effective solutions.

Broken into its constituent steps, the cause mapping process functions as follows:

0. Divide the world up into primary causes

1. ‘Screen out’ primary causes where it’s clear all their solutions are cost-ineffective. This is done by appealing to one or more of scale, neglectedness, and solvability individually.
2. Of the primary causes that remain, list the main mechanisms available for each primary cause.
3. Assess the main obstacles in the way of each mechanism.
4. Create a list of solutions by combining 2. and 3.
5. From the list of solutions, set out secondary and meta-causes.
6. Evaluate the solutions for cost-effectiveness.

Readers may wonder how this is different from the EA method, either as originally stated or on my reconceptualisation. There are two comments to make.

First, I do not think it’s the case the above steps are incompatible with, or different, from those in the EA method. Rather, I am simply making explicit the different steps that were already implicit in EA method (and that one must undertake when thinking about how to do the most good). What occurs in steps 2-4 of cause mapping is something that must implicitly occur *between* steps 1 and 2 of the reconceptualised EA method—to get to step 2 of the latter, we must acquire a set of particular solutions from somewhere. Cause mapping is about filling out the items on that set.

Second, the EA method, at least as articulated by MacAskill and others, does not encourage or require us to map out the different mechanisms, obstacles, solutions and so on.[3] While this is not a radical innovation, it does seem useful to carefully work through the different steps in the hope of discovering altruistic opportunities that were not *prima facie* obvious—this is, indeed, what I found when applying it to the question

[3] See the long MacAskill quote at the start of chapter 5.

that follows. Given there seems to be no *a priori* means of working out what our altruistic priorities are in the actual world, this sort of approach seems to be the best we can do.

Thus, cause mapping can be seen as an attempt to break down the EA method into its smallest distinct steps, record the most promising items considered, and identify how those items connect together to produce a reasonably comprehensive list of altruistic options. Once that list is in hand, the subsequent step is to evaluate those options for cost-effectiveness.

2. Using cause mapping to find ways to make people happier

In this section, I apply the cause mapping method to the question of how best to improve lives, where improving lives refers to increasing people's well-being while they are alive. As in chapter four, I assume well-being consists in happiness. To explain the focus on improving lives, as opposed to any other problem, e.g. reducing extinction risks to humanity, I am sympathetic to Person-Affecting Views on population ethics (on which there is no value in creating new lives) and to hedonistic utilitarianism (the right act is the one that maximises happiness).[4] On this combination of views, we ought, quite literally—to paraphrase Narveson's Dictum— to make people happy rather than make happy people.[5] The two potential ways to increase someone's lifetime well-being are by improving or lengthening their lives, i.e. enhancing quality or quantity of life. Chapters 1 to 3 discussed the latter at length. Chapter 4 argued we can

---

[4] For a general discussions of population ethics, see (Greaves, 2017) and (Arrhenius, unpublished). For arguments on person-affecting view, in particular, see (Bader, no date) and (Heyd, 2009, 2014) For a classical argument for classical utilitarianism, see (Mill, 1861) and (Bentham, 1789).

[5] (Narveson, 1973). I assume in his expression 'morality is in favour of making people happy and neutral about making happy people', Narveson used 'happy' as a placeholder for well-being and his statement is not necessarily an endorsement of hedonism (well-being consists in happiness).

measure happiness and discussed how, assuming we want to make people happier during their lives, moving to happiness-based cost-effectiveness might alter our priorities from those presently suggested by effective altruists. However, that left many possible ways to improve lives unexamined; this mapping aims to fill that gap. I do not defend person-affecting views or hedonistic utilitarianism here, nor does it seem necessary to do so—nearly everyone holds making people happier has some value and will want to know how best to do it and how it compares to whatever their current altruistic priorities are.

For definiteness, I approach this question from the perspective of a hypothetical philanthropist looking to spend their money on making other people happier. The other obvious perspectives to take would be to consider what the most impactful careers are for altruistic agents to take or what governments could do. It's harder to evaluate the cost-effectiveness of time put into different careers than money put into different organisations—people have different skills. Thinking about government policies raises different issues that are too complicated to consider here.[6] Hence, I stick to the simpler, pecuniary question facing private agents. Further, while I will concentrate on improving the lives of currently existing persons, although almost all the analysis could be reused if one wanted to also consider how to make all possible people happier during their lives.[7] I note that my investigation here is necessarily highly limited: investigating this question could be at least an entire thesis in itself. As

[6] Some problems: do we have a developing or a developed country government in mind – presumably the policy prescriptions would be different for e.g. the UK vs Zimbabwe? If we're talking about taking money from one area, e.g. defence, and putting it into another, e.g. health, it becomes hard to calculate the expected value of those trade-offs – e.g. the value of an extra peace-keeping mission vs X person being treated for mental health. Recent work on happiness-based public policy has mostly on individual policy areas without discussing trade-offs, see e.g. (Sachs *et al.*, 2019)

[7] Peter Singer helpfully notes that someone concerned with the welfare of all possible lives would also need to consider other questions, for example regarding how many people to bring into existence.

such, to borrow an analogy from actual mapping-making, my aim in this chapter is similar to that of Magellan, the first person to sail around the world: I want to make an initial 'map' of how things fit together while recognising this map cannot be very detailed or accurate and highlight where I have not be able to go.

Cause mapping for this domain is presented as follows. Section 3 discusses the priority primary causes. Section 4 considers, in hypothetical terms, what the available types of mechanisms and obstacles are. Section 5 presents the most relevant mechanisms and obstacles, given the priority primary causes. Section 6 presents a list of secondary causes. I will note in advance that my discussion does not stretch to a detailed consideration of meta-causes—more on this is said in section 6. Section 7 sets out what further work is required and makes some concluding remarks.

3. What are the priority primary causes?

The first question to ask is: what are the problems we want to address so as to improve lives as much as possible? As improving lives refers to increasing people's happiness during their lives, before we can answer this question, we need to say what we mean by happiness and how we measure it. Here, I restate the answers given in chapter 4, which is that happiness is the positive balance of pleasure over pain. While the most accurate way to measure how happy people are, given current technology, is to track individuals' experiences and ask them to report how good/bad they feel at randomised moments, this is not very practical to do and there is resultantly not that much data available.[8] As such, I suggested the most practically suitable measure for happiness is

[8] How good/bad they feel with respect to their levels of pleasure and pain (or displeasure).

life satisfaction, which is normally found by asking people 'Overall, how satisfied are you with your life, nowadays?' (0—10).

We can now turn back to the question. One way to answer it would simply be to list absolutely everything that, if we addressed it, would take us closer to maximum possible happiness—wars, climate change, fickle politicians, noisy neighbours, poor journalism, Mondays, etc.—but not only would this list be impossibly unwieldy and hard to manage, it would also not consist solely in primary causes. Fixing problems, such as poor journalism and bad politicians, would not, in themselves, increase happiness, but would do so indirectly through other things—the better politicians would enact better policies that would increase happiness. Those are what I dub 'meta-causes' for that reason. We must start with the primary causes and return to meta-causes later.

As mentioned in the previous section, to evaluate what the priority problems are I follow the first step in the reconceptualised EA method: screening out problems that seem to have no solutions better than the best ones we are already aware of. Because the current top life-improving interventions, as proposed by effective altruists, involve alleviating the poverty of the world's poorest people, I will use those as the point of intuitive comparison: if a primary cause doesn't seem to have a solution that is more cost-effective than those interventions, I leave it off the list. I note in advance this analysis relies on the author's subjective judgements; while this is perhaps unsatisfying, the relevant alternative—to produce evidence-based, quantitative assessment of all, or even some, of the different solutions before declaring them to be cost-ineffective—is impractical.

I'll state the four primary problems that seemed most promising and why that is the case, then explain what was also considered and why those alternatives didn't make

the list. The four problems are poverty, mental health, pain and what I'll call 'mundane, sub-optimal happiness', hereafter shortened to 'MSH'. It is reasonably clear what the first three refer to, so I'll only explain the fourth. The idea behind MSH is that it captures the gap between maximum possible happiness and the level of happiness people experience when living what we could call 'a fortunate life'—they are mentally and physically healthy, their material needs met and not experiencing any obvious major 'life event', such as bereavement, divorce, crime, bankruptcy, marriage, unemployment, becoming a parent, etc. The focus on MSH directs our attention towards the sub-maximal experiences in the mundane, unexceptional moments of our lives. At this level of analysis, it is unnecessary to define any of the primary causes more precisely.

Let me say something fairly general about why those four causes look promising. It seems the key results from the empirical research on happiness are as follows. First, 'hedonic adaptation', we get used to most things over time, so that few things make us feel very good or very bad for very long.[9] Second, 'social comparison', we often judge ourselves against others so that one person's gain can be another's pain. [10] Third, failures of 'affective forecasting', ('affect' is a psychologist's term for emotion), that is, we struggle to accurately predict how we'll feel in the future; the general rule is we overestimate the duration and magnitude of good/bad events.[11] Evolutionary theory provides an explanation for these results: happiness is 'nature's' reward and punishment system that steers us towards activities that aid survival and

[9] For a general theoretical discussion of adaptation, see (Frederick and Loewenstein, 1999) and (Diener, Lucas and Scollon, 2009). For empirical evidence of adaptation to specific events (Luhmann *et al.*, 2012) and (Clark *et al.*, 2018).
[10] See e.g. (Alderson and Katz-Gerro, 2016), (Kim *et al.*, 2017), (Clark, 2017).
[11] See (Ayton, Pott and Elwakili, 2007), (Wilson and Gilbert, 2005), (Morewedge and Buechel, 2013).

reproduction.[12] Thus, for instance, we shouldn't expect to adapt—get used to—to sensations like pain, hunger or sadness because their function is to keep us feeling bad so we're motivated to act.

Focusing on non-adaptive and non-comparative (i.e. not effectively zero-sum) problems seems like a good place to start, as all others will either have a short-term effect or not do much to increase overall happiness.[13] The four primary causes are the obvious examples of non-adaptive, non-comparative problems where we are concerned with improving lives.[14] Specifically, that is why I focus on pain rather than physical illness: there are many physical illnesses but the thing that is distinctively bad about them, which is also non-adaptive and non-comparative, is being in pain, hence that seems to be the thing to target.[15] Regarding MSH, the idea that it is possible to improve our everyday experiences can be regarded as a hypothesis that I later suggest is plausible.

To see what else might have been on the priority list, we (re)turn to figure 6.1, which featured in chapter 4; this lists the effects of different life changes in life satisfaction.[16] As noted there, this analysis is state of the art, using a national panel data set (the same

---

[12] For discussions from an evolutionary theoretical perspective see (Ng, 1995), (Rayo and Becker, 2007), (Perez-Truglia, 2012), (Ahuvia, 2008).
[13] In effect, I am appealing to scale here. As discussed in the previous chapter, if a problem is very small, e.g. helping a particular bee, we can quickly intuitive it won't be the most cost-effective use of our resources.
[14] I am grateful to Peter Singer for observing that, if we were concerned with improving and saving lives, then having's one life shortened would be a non-adaptive, non-comparative problem: you do not 'get used' to being dead, nor do we seem to be made happier if we learn other will live less long; on the contrary, others living longer is a taken to be a good sign as we infer we will live longer.
[15] It's hard to believe one person's experience of pain makes others happier by an equal and opposite amount. What's more, as (Dolan and Metcalfe, 2012) show, using subjective well-being data (both affect and life satisfaction) is that it is the mental distress and pain components of poor health that have the bigger impacts on subjective well-being, as opposed to those related to increased difficulties in mobility, self-care or performing usual activities (the other 3 components of the EQ-5D, a health questionnaire used to determine QALY-weights).
[16] (Clark *et al.*, 2018) at p199.

people were surveyed each year), allowing individuals to be used as their own controls and changes observed over time.

	Effect on life-satisfaction (0–10)	*Total effect on the life-satisfaction (0–10) of others*
Income doubles	**+0.12**	**–0.13**
One extra year of education (direct effect)	**+0.03**	**–0.09**
Unemployed (vs. employed)	**–0.70**	**–2.00**
Quality of work (1 SD extra)	**+0.40**	—
Partnered (vs. single)	**+0.59**	**+0.68**
Separated (vs. partnered)	**–0.74**	—
Widowed (vs. partnered)	**–0.48**	—
Being a parent	+0.03	—
One physical illness	**–0.22**	—
Depression or anxiety	**–0.72**	—
Commit one crime	**–0.30 point-years**	**–1.00 point-year**

Figure 6.1

Given that we are thinking from the perspective of a philanthropist, rather than a government, it seems hard to think of what could cost-effectively be done about unemployment, quality of work, being partnered, being separated, being a parent, or committing a crime. To emphasise one of these, even though being partnered has a huge impact on life satisfaction, it difficult to see what a well-meaning philanthropist (or, in fact, government) could cost-effectively do with their resources that would improve on individuals' efforts to find love.[17]

If we strike out those life changes, that leaves the effects of income, education, physical illness, and depression or anxiety. I further rule out education, in the sense of providing additional, conventional academic education, because (a) this seems not to

[17] This is not to say there is nothing that could be done. The philanthropist could invest in research into 'love drugs', as (Earp and Savulescu, 2018) suggest, or provide couples counselling.

have large positive effects for recipients, and (b) there seems to a negative social-comparison effect on non-recipients when others are more educated (see figure 6.1).[18] I don't rule out the effects on income, despite its other-regarding effect, because this is data from the developed world and it is plausible, as discussed in chapter four, that raising the income of the very poor does increase aggregate happiness.

Besides the above, other events that we would expect to have a long-term (i.e. non-adaptive) impact that I considered, but didn't seem particularly tractable, were being imprisoned, enslaved, lonely, experiencing war, and the effects of climate change. For all but the last, it is also hard to think what interventions a philanthropist could take that would be anywhere near as cost-effective targeting global poverty. I did not include climate change as modelling what its impact on happiness would be is too complex to take on here (e.g. what would we expect the effect of sea rises to be on happiness and how much would preventing the release of X tonnes of carbon alter this by?).[19]

We could also think about causes which would increase happiness over the short-term—as Keynes noted, 'in the long run, we're all dead'[20]—and these could potentially still be cost-effective to do things about them. Etile et al. provide a handy set of 10 life events and show that people eventually adapt to all of them.[21] These are: (1) major

---

[18] The explanation of this is that education functions as a positional good, much like money – it is better for me to more educated, but this comes at a cost to you. It is puzzling that figure 6.1 shows the effects of extra education of others is negative and (three times!) larger than benefits to those who receive more education. I have asked Richard Layard, one of the cited book's authors about this (personal conversation): he was not sure why the negative, other-regarding effect should be this large.

[19] While there are estimates of the economic costs of climate change, it's not straightforward to think what impact such effects will have on happiness. The worry here is the evidence from the Easterlin Paradox, where rising incomes do not seem to increase aggregate life satisfaction. Hence, it's an open question how much reducing economic activity will reduce life satisfaction and thus happiness. See (Stevenson and Wolfers, 2008) for a critique of the paradox's existence and (Easterlin, 2016) for a rebuttal.

20 (Keynes, 1923) p. 80

[21] (Etilé *et al.*, 2017)

financial worsening; (2) fired or made redundant; (3) separation from spouse; (4) death of spouse or child; (5) death of close relative; (6) death of close friend; (7) injury or illness to self; (8) injury or illness to relative; (9) victim of physical violence; (10) victim of property crime. I exclude these as unpromising for various reasons. For (1)—(3) and (9) and (10) it is unclear what a philanthropist could usefully do. (4)—(6) relate to death and the value of saving lives, which I have discussed elsewhere. (7) and (8) would potentially be captured by pain and mental health treatments anyway.

I also considered more exotic problems, such as a global pandemic, nuclear war, rogue artificial intelligence or an asteroid strike. I assume their importance is based on their impact on reducing the number of future lives, rather than on their impact on the quality of people's lives (either in the near or long-term).[22]

While this set of problems covered here is reasonably comprehensive, I do not claim it is exhaustive. There is also the serious possibility that there are cost-effective solutions to some of these problems that I have mentioned above.

4. What types of mechanisms are available? What types of obstacles block those mechanisms? (Theoretical)

Now we have the primary causes we wish to solve we need to look for some solutions. My earlier suggestion is to split solutions, particularly the actions we can take, into two component parts: mechanisms and obstacles. In this section, I'll explain each of the parts in turn and propose typologies of them. In the next section, I will apply these typologies to the primary causes.

[22] Although I note in chapter 3.3.1 I discussed (Lewis, 2018) who makes a case that reducing existential risk is still a 'good buy' compared to other health spending, if one wished to save lives.

The first component of solutions are the happiness *mechanisms*, the different ways that we can change parts of someone's life to try to increase their happiness. Schematically, it seems we can split these into six categories, for which I provide an illustrative example in each case:

1. External: altering someone's objective circumstances, such as wealth, education, physical environment or the society they live in.
2. Temporal: how people choose to spend their time.
3. Psychological: changing how people think, e.g. cognitive treatments for mental health.
4. Chemical: using mood-enhancing substances, e.g. alcohol, painkillers or anti-depressants.
5. Physical: direct manipulation of the body or brain, e.g. surgical procedures, Deep Brain Stimulation ('DBS').
6. Biological: genetically modifying people to be happier.

These are all meant as types of mechanisms we can use at least once for any of the primary causes. It is not obvious that all six mechanisms can be used for every cause.

The motivation for creating *a* categorisation is that it gives us a toolkit, or a check-list, that we can refer to when thinking about finding the mechanisms for any happiness-related cause. The motivation for using *this particular* categorisation is these seem to be the full range of different pathways we can use to eventually change someone's conscious experiences.[23] Of course, having a typology doesn't guarantee we'll find every possible item of each type, but it does at least prompt us to look.

[23] This is not the only possible typology. One alternative would be to have a three-way split into external, temporal and internal, where 3 to 6 on the list were lumped in with internal. Another would be sub-divide external further, possibly into capital (both human and economic), environmental and social.

I noted earlier that the analysis would not change much if one considered improving just the lives current people or of all possible people. The case where it would likely change relates to the biological mechanisms. Arguably, altering the genetics of a foetus would be tantamount to preventing one person from existing and causing another to exist. On a *narrow(/strict)* Person-Affecting View in population ethics, where all that matters is how things go for particular individuals, there would be no value in making this change: it is not better for either the pre-modified or the post-modified person to be created and have a happy life, and hence the change is not better overall (ignoring the effects on other people, if any).[24] However, as Savulescu and Kahane argue, on a *wide* person-affecting view, it would be better to engage in gene-editing in cases where we value the well-being of the people that come into existence, whomever they happen to be, because it would be better for *people* (although not for the particular persons).[25] On a wide person-affecting view, gene-editing seems a very promising way to improve lives.

Obstacles are the second component of solutions. Once we've found a particular mechanism that helps with a particular primary cause, we can ask 'What is stopping that mechanism from being used to benefit as many people as possible'? Our answer to that question indicates what we think the obstacles are. In general terms, there seem to be five types of obstacles:

- Research: the mechanism is (theoretically) usable, but more know-how is required

[24] This, indeed, is Parfit's famous Non-Identity problem. See (Parfit, 1984) Chapter 16. For distinctions between different Person-Affecting Views, see chapter 3.3, footnote 26.
[25] (Savulescu and Kahane, 2009)

- Behaviour/motivation: people don't want to use the mechanism, even though it's available
- Education: people would use it, but they don't know about it
- Resources: people know about it and would use it but can't afford it
- Policy: the state needs to act before the mechanism can be used

For each mechanism, we would expect all or nearly all the obstacles to apply to some extent, so the particular question to ask is: which obstacle(s) should we target if we want to have the most cost-effective impact? For instance, one way of increasing happiness would be providing psychotherapy to people with mental illnesses. Intuitively, the obstacle seems to be money. If someone came along and funded more services somewhere in the world, presumably additional people would use them. However, this might not be the case, or might not be the case everywhere—in some places, perhaps the bigger obstacle is stigma, such that people are not motivated to take up services even if they are available. Ultimately, we would want to rely on empirical research to investigate which obstacles are more significant.

It is important to note that the question 'which is the most important obstacle?' can only usefully and sensibly be asked from the perspective of a particular agent, rather than from the perspective, as it were, of the world. Different types of agents will have different capacities, so we need an agent in mind. For instance, it might be more effective for an influential celebrity to try and educate people about mental health, a doctor to do research, a politician to campaign for policy change, and a hypothetical billionaire to fund an organisation providing mental health services. When we ask, as we often do in relation to some problem, 'what needs to be done?' we must, at least implicitly, have some (sort of) agents in mind who we expect should act. I think in pecuniary terms in this analysis as the analysis is the same whoever's money it is. As a

methodological aside, if we wanted to redo this analysis with a different agent in mind, we would only need to recalibrate it from the point of obstacles onwards; everything before that seems to be agent-neutral.

5. What types of mechanisms are available? What types of obstacles block those mechanisms? (Applied)

The previous section focused on setting up the parameters that are relevant for determining the solutions in general, theoretical terms; however, we've not yet filled in the details. This is what happens next. As there are four causes and six categories of mechanism, that gives us twenty-four different potential sets of solutions.

I've put my results from this process in table 6.1, the 4 by 6 box that follows in two pages. What I'll now do is quickly describe the contents of the parts of the box in turn, stating the particular mechanisms and what seems to be the relevant obstacle in each case. While there are 24 individual boxes to cover, we can avoid a repetitive analysis by grouping the discussion by mechanism type. I note that, as we are engaged in cause mapping, the objective is only to list the main possibilities. Making an exhaustive list is unfeasible. Analysing the possibilities in detail and evaluating their comparative cost-effectiveness is a further empirical task subsequent to the cause mapping process. I note my concern here is solely with trying to understand what might increase happiness, not with the ethics of using this or that intervention—the moral question is distinct from the one I pursue and an area for further work.[26]

[26] For, e.g. some discussion of the ethics of human enhancement, see (Savulescu, Muelen, and Kahane 2011).

5.1 External

The external mechanism captures ways we could change the world around people, as opposed to changing them in some way. External is a reasonably capacious category and can be broken down further into capital (i.e. resources), social, and environmental changes.

Starting with poverty, to simplify matters, I assume that poverty is only solvable by the provision of extra resources—if we can make those in poverty happier by (say) changing how they think then I stipulate that to be part of MSH instead. Following GiveWell, the particularly promising interventions for poverty are providing cash transfers or deworming, and the obstacle to more of these happening is simply money.[27] Also plausible are systemic changes, such as attempting to reform international trade laws or immigration policies, where the obstacle is policy.[28]

Moving on to social and environmental external changes, we could try to improve MSH by making people friendlier, more moral, less corrupt, more trustworthy, reducing the inequalities with societies as well and making physical spaces more pleasant and less polluted, to name the main options.[29] The relevant obstacles here seem to be policy (e.g. for corruption, inequality, and pollution) as well as behaviour (for the others). The philanthropist could fund organisations that spread prosocial attitudes—building the effective altruist movement being an obvious example—or advocate for various

---

[27] (GiveWell, no date b)
[28] E.g. (Wenar, 2015) sets out how current international trade rules cause poverty.
[29] See (Helliwell, Layard and Sachs, 2019) ch.2 on how international differences in subjective well-being can by six key variables: GDP per capita, social support, healthy life expectancy, freedom to make choices, generosity, and perceptions of corruption. On green space, see e.g. (Bertram and Rehdanz, 2015).

changes at the policy level. If it were possible to teach such kindness and considerateness in schools, that would seem an obvious place to do so.

In the external category, pain and mental health would also be the social determinants of those conditions: what the WHO calls the situations in which "people are born, grow, live, work and age".[30] These are very diffuse, encompassing employment, social exclusion, globalisation, gender, and so on. They are sufficiently diffuse that I do not investigate them here—it seems unpromising to focus on any one of these many contributing factors rather than directly addressing the problems at hand.

[30] (WHO, 2017)

	1. External (capital; environmental; social)	2. Temporal	3. Psychological	4. Chemical	5. Physical	6. Biological
Mundane, sub-optimal happiness (MSH)	Society and culture change (behaviour; policy); physical environment (money; policy)	Lifestyle changes, including 'nudges', e.g. more exercise, shorter commutes (behaviour; education)	Thinking training, e.g. mindfulness, positive psychology, resilience training (education; behaviour)	Recreational drug use (policy; research)	'Wireheading', i.e. using wires to directly stimulate the brain (research; policy)	'Hedonic enhancement', i.e. genetically modifying people (research; policy)
Mental health	Social determinants of mental illness (behaviour; education)	Lifestyle changes, including 'nudges', e.g. more exercise, shorter commutes (behaviour; education)	Talking therapies, e.g. cognitive behavioural therapy (money; education)	Anti-depressants (money; regulation); psychedelic-assisted therapy (research; policy)	Electrical treatments (e.g. Deep brain stimulation, transcranial magnetic stimulation) (research)	Hedonic enhancement (research; policy)
Pain	Social determinants of health (behaviour; education)	N/A	Cognitive pain strategies, e.g. mindfulness (education; money)	Access to opiates in the developing world (policy; research)	Electrical treatments (see above) (research)	Hedonic enhancement (research; policy)
Poverty	Development projects (money); international reform (policy)	N/A	N/A	N/A	N/A	N/A

Table 6.1 A mechanism and obstacle matrix for the four priority primary causes.

5.2 Temporal

As people enjoy some things more than others, a natural thought is to nudge people into making changes to their lifestyle, so they do more of the things they do like, e.g. socialising and exercising, and less of the things they don't, e.g. commuting and working.[31] This most clearly improves MSH, but would also improve mental health: for instance, one treatment for mental health is Behavioural Activation where people are encouraged to be active and do things they enjoy.[32]

The interventions we're considering here are those which individuals are already free to undertake. Presumably the obstacle is education and motivation: similarly, there are already huge public efforts to encourage people to engage in healthier behaviours, such as smoking and drinking less and exercising more, and these are often tackled through campaigns, as well as policy changes (for instance, making cigarettes more expensive). Encouraging people to change their time use is something that could and would be included in any 'positive education' curriculum, which I will come back to in the next section.

5.3 Psychological

Teaching people to change how they think works for mental illness, pain and MSH. This is a major way of treating mental illness, where Cognitive Behavioural Therapy, which teaching people to understand and change their thinking patterns, is a standard and effective approach;[33] a more recent approach is mindfulness, a secular form of meditation.[34] The main obstacle here seems to be money: people would use such

---

[31] For an example of a time-use study indicating this, see (Kahneman *et al.*, 2004)
[32] (Ekers *et al.*, 2014)
[33] (Cuijpers et al., 2013)
[34] (Khoury et al., 2013)

treatments if they were available for free. That said, many people who could afford to see a therapist privately—or treat themselves using freely available resources, e.g. on the internet—seem not to do so, which raises the question of whether education and/or motivation are the obstacles.

Regarding pain, there's evidence that cognitive strategies can be used to treat this.[35] The obstacles here seem similar to those for mental health.

It turns out the same types of interventions, e.g. cognitive behavioural therapy and mindfulness work on non-clinical (i.e. healthy) populations, which means they could be used for MSH too.[36] One promising idea to get this to people is to teach various thinking and resilience techniques in schools. This is approach is called 'positive education' and research testing it in very large populations (i.e. in the hundreds of thousands of students) indicates it raises self-reported well-being scores, as well as standardised academic test scores.[37] Note that positive education—teaching and providing psycho-social skills—is different from 'conventional' education, the imparting of academic knowledge about, e.g. biology and history. Hence, while providing extra positive or conventional education may have a positional effect (see section 3), the former has a greater positive impact on the subjective well-being of recipients. Another avenue for spreading this useful information would be to run public well-being campaigns, which are either directly funded by the philanthropist or use private resources to lobby governments to do so.

---

[35] (Ehde, Dillworth and Turner, 2014)
[36] See e.g. (Cukrowicz and Joiner, 2007; Kaviani, 2011)
[37] See e.g. (Sachs *et al.*, 2019) chapter 4 on positive education.

I haven't included any temporal mechanisms to alleviate pain as it's unclear what they could be.

5.4 Chemical

Various chemical interventions could help with the primary causes. If we start on mental health, it is clear anti-depressants are a widespread and somewhat effective treatment. In the more developed world, these seem to be relatively cheap and easily available from doctors, hence the obstacle—assuming more of them could be more useful—is presumably that people do not choose to ask for them (behaviour). In the developing world, access to anti-depressants (like many medicines) can be hard to come by. Presumably, the obstacle here is resources.

There's evidence that some currently illegal drugs, such as LSD and psilocybin, show great promise in the treatment of mental health conditions.[38] Their use, even for research, is typically heavily restricted by government regulation, hence policy seems the obstacle.[39]

Turning to pain, a recent commission by *The Lancet* noted a lack of access to opiate-strength painkillers in the developing world, which means many people there suffer avoidably (this should be distinguished from the problem of opiate 'over-prescription' that currently affects America). The issue here is again policy—developing world governments would need to improve access.

Regarding MSH, there's a possibility some form of drug liberalisation may improve happiness.[40] There are several potential reasons for this here, which are: casual

---

[38] (Nichols, Johnson and Nichols, 2017)
[39] (Nutt, King and Nichols, 2013)
[40] Depenalisation, decriminalisation and legalisation are all different options for liberalisation.

recreational use may improve lives; liberalisation could reduce harm to users; it could reduce crime and increase development around the world—countries like Mexico and Columbia are hamstrung by drug conflicts. The obstacle to change is policy here too.

5.5 Physical

While increasing happiness through physical mechanisms, such as direct brain stimulation, may seem very 'sci-fi', I note this already occurs: deep brain stimulation (DBS) has been used to treat pain and depression;[41] repetitive transcranial magnetic stimulation (rTMS) has been used to treat depression.[42] Technologies such as these can already be used by consumers.[43] They could potentially increase MSH, but it remains to be seen whether they do or not.[44] The obstacles here seem to be research—i.e. more know-how is required to investigate and increase their effectiveness—and also policy issues related to their use in medical and consumer contexts.

5.6 Biological

As noted earlier, genetically modifying people to become happier is at least theoretically possible.[45] Just as for physical mechanisms, the relevant obstacles appear to be research and policy.

6. Which interventions share the same obstacles? What secondary causes do we get as a result?

The previous sections identified different mechanisms by which we might improve lives and the obstacles in each case. What we can do now is group together cases where

---

[41] See e.g. (Kennedy *et al.*, 2011)
[42] (Ren *et al.*, 2014), (George *et al.*, 1995)
[43] (Maslen et al., 2013)
[44] (Wexler and Reiner, 2019), (Page, 2001)
[45] (Savulescu and Kahane, 2009)

the obstacles are all the same to form *secondary causes*, which are the different candidate solutions for making progress on the *primary causes*. I propose six secondary causes. These turn out to be substantially, but not entirely, the same as the groups of interventions discussed in the previous section. This is perhaps not surprising—for instance, if we're thinking about changing how people think, we would expect to run into many of the same issues, whether we are trying to alter their thinking to help them with mental illness or with pain. Similarly, for the chemical mechanisms, a broad, shared obstacle is policy—access to such substances is widely regulated.

I'll now list the six secondary causes, providing the subsection of section 5 from which they are drawn. Where it's obvious what the secondary cause is, I will not provide further commentary in order to avoid repetitiveness. How the secondary causes emerge is indicated in the graphical representation of the cause mapping in figure 6.2. After setting these out, I make some brief remarks on meta-causes.

1. Poverty alleviation (5.1).
2. Positive education (5.2 and 5.3). This focus on helping people to improve how they think (a psychological mechanism) and how they spend their time (a temporal mechanism). Note this secondary cause would also serve the purpose of building resilience to all the short-term 'life-events' I mentioned in section 3 but didn't include as primary causes in their own right, such as losing a partner, being a victim of crime, becoming unemployed, etc.
3. Improving access to psychological therapies for mental health (5.3).
4. Drug policy reform (5.4). In the chemical mechanism section, it was clear that, both in a medical and recreational context, quite a few specific ways of increasing happiness would be available if drug policy was changed. Hence a broad secondary cause is drug policy reform.

5. Social, cultural and environmental change (5.1). This picks the social and environmental changes in the external mechanism sub-section.
6. Happiness, mental health and pain research (5.4, 5.5, 5.6). While lack of insight is frequently a problem, it appears to be the main obstacle for the last three mechanisms we have discussed so these have been grouped together

That completes the list of six secondary causes.

In figure 6.2, I have also included three meta-causes. I don't propose to say much about meta-causes because, in order to evaluate the cost-effectiveness of meta-causes, we need to be able to tell a story about how resources to the meta-causes will eventually translate into better outcomes among the primary causes through affecting the secondary causes anyway ('how exactly will spreading altruistic values increase donations to charities? By how much will it do this and where will the new money go? Etc.'). Indeed, a test to identify whether something is a meta or a secondary cause is that you can't calculate the expected value of the former unless you already know the expected value of additional resources in the latter. Seeing as we currently can't say much about the efficacy of the secondary causes here, it seems we should make more progress on those before returning to this topic. It is worth saying that if we want to assess the impact of meta-causes accurately, we'll need to consider what sort of impact they will have on everything *else*. This is something made much more straightforward by having done the cause mapping approach. The arrows I've drawn from the meta to the secondary causes are merely illustrative and indicate the direction of causality, in the sense that meta-causes are impactful via their effects on secondary causes.

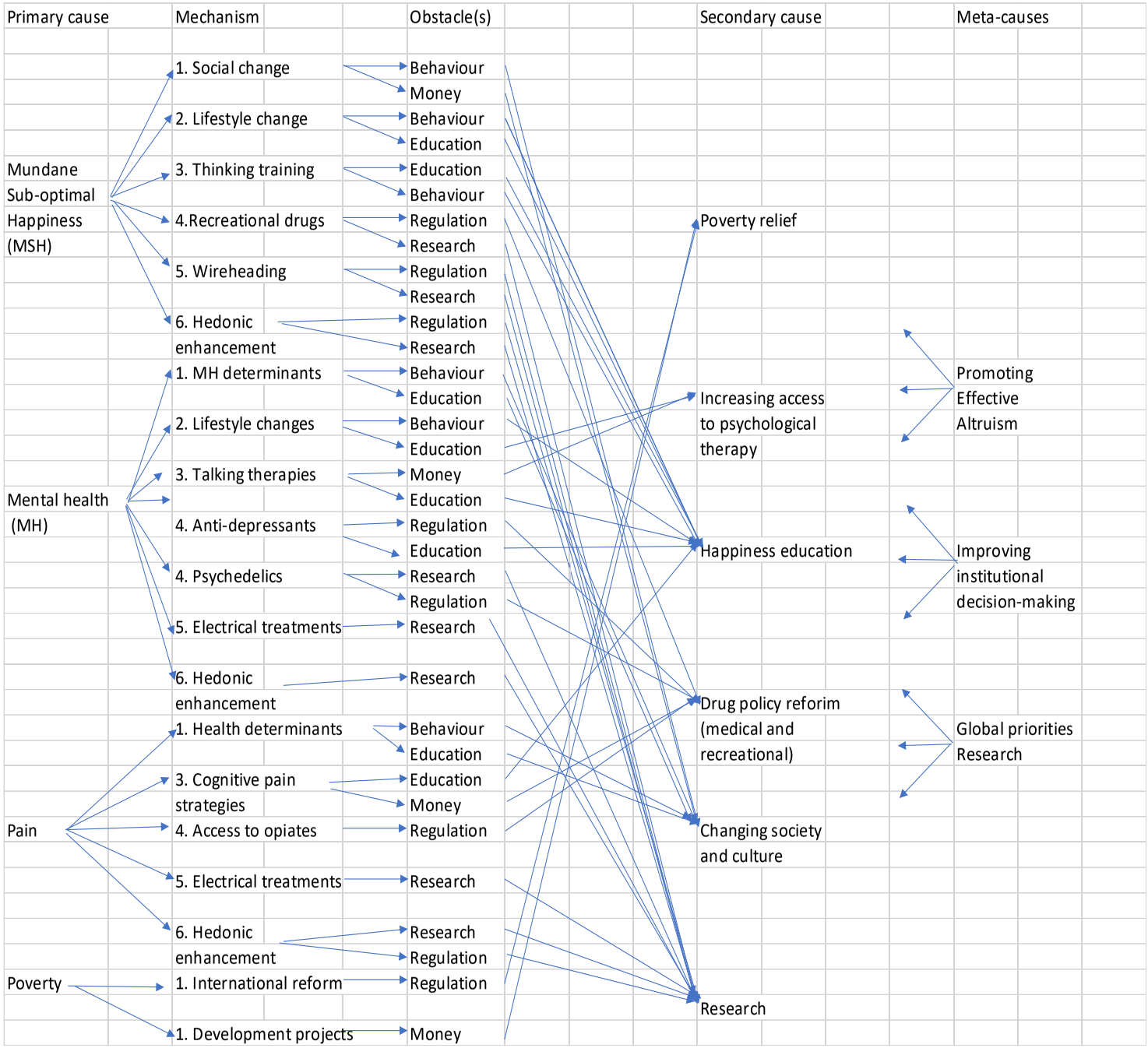

Figure 6.2. Causing mapping for maximally improving lives. Notes: for mechanisms, these are preceded by a number indicating which type of mechanism they use; the arrows between the Obstacle and Secondary cause columns show how the obstacles collect into secondary cause; the arrows from meta causes are indicative only.

7. Conclusion

In this final section, I tie up a few loose ends: clarifying the nature of the secondary causes and how to make progress on them; the potential oddness that my secondary causes are at odds with the current suggestions of effective altruists; the usefulness of cause mapping in other domains.

The secondary causes are best understood as a 'longlist' of areas for further investigation in the sense that we hope they will contain all the good options, but not all the options they contain are good—by differentiation, we would say a 'shortlist' would have only good options. As such, more work is required to carefully look through them. This would start with explicit, quantitative expected value calculations of particular options, from which some conclusions could be reached about what the best ways to improve happiness.[46]

More specifically, making progress on evaluating the secondary causes requires doing one or more of three tasks. First, running new empirical experiments to test key claims. For instance, running randomised-controlled trials using happiness measures to test the cost-effectiveness of mental health and anti-poverty interventions in the developing world. Second, building cost-effectiveness models of interventions where evidence already exists, such as on the effect of positive psychology programmes that teach children to be happier. Third, establishing the particular places money could be donated to and comparing the effectiveness of particular organisations in that field. For instance, it is unclear how best to pursue drug policy reform: would it be a ballot initiative in California to legalise all drugs, funding research into the using of

[46] A more detailed set of specific questions for which research would need to be done or, where it exists, collected, can be found here (Plant, 2019).

psychedelics to treatment mental health conditions, or something else? Even having considered that, organisations in that particular domain require assessment.

Clearly, there is a huge amount of further work to be done here, much of it outside the domain of academic philosophy. While I am unable to complete all this analysis, I will attempt to take a bite out of it in the final chapter, 7, where I compare charities that alleviate poverty—which are currently favoured by effective altruists—to one that treats mental health, and do so in terms of subjective well-being scores. Having investigated prioritisation methodology in this and the previous chapter and then, in this one, looked through the possible options, mental health interventions—specifically, mental health treated through psychotherapy—stands out as the first option to investigate if we are looking to find something more cost-effective than alleviating poverty.

This is for a couple of reasons. First, treating mental health via psychotherapy is, intuitively, among the most cost-effective options we discovered, particularly given (as noted) the large per-person badness of mental illness is relative to other life-changes and the existence of effective treatments. Second, it allows a direct comparison: we can compare the efficiency of developing world charities that directly provide evidence-based interventions of one type—poverty alleviation—against one that provides another—treating mental illness. Third, it would be more striking if—as a result of using different approaches to measuring happiness and priority-setting—some novel problem turned to be more cost-effective than the current priorities of some known problem.

Before we move to this analysis in the next chapter, there are two further points to make. As noted in chapter 4, effective altruism, in as much as it is concerned with benefiting those alive today, seems to focus on poverty and physical health. That might

be taken as evidence I've made some serious error, given that the secondary causes listed suggest there are many additional options worth investigating. The explanation here, as mentioned before, is that effective altruists seem to not have seriously considered how best to improve lives and/or have not incorporated the empirical evidence on subjective well-being in their analyses. Certainly, no one has explicitly made claims about which are the best things to do if one wants to improve lives and provided arguments or evidence I can critique. Hence, my suggestions could be understood as a reasonable list of options that emerge when this question is investigated seriously for the first time.

The final point is about whether cause mapping should be used in other areas besides maximally improving lives, such as when the question is how to maximise good over the long-term. This seems at least worth trying: it may clarify thinking and indicate new topics to be investigated but there is no guarantee of this. To re-run it in other domains would require a new list of priority primary causes and a different typology of interventions, but the rest of the approach would remain unchanged and may prove to be illuminating.

Chapter 7: Spending money, buying happiness

0. Abstract

This chapter asks: if one wanted to give to charity to make people happier, and do so as cost-effectively as possible, is alleviating poverty the priority? The answer to this question among effective altruists, notably Singer and MacAskill, seems to be Yes. Against this, I argue that, if we measure happiness using self-reported subjective well-being scores, treating mental health looks better.

1. Using subjective well-being score to estimate charity effectiveness

The current view among effective altruists, as exemplified by Singer and MacAskill, is that global poverty is the priority if we are trying to make people alive today happier.[1] In chapter 4, I noted that neither Singer or MacAskill make use of the social science literature on self-reported subjective well-being (SWB) and then argued SWB scores can and should be used to measure happiness. In this chapter, I attempt a first-pass of what the most cost-effective charities are if we evaluate them through the 'happiness lens', that is using SWB scores as the measure of success. Specifically, I use life satisfaction scores and then convert from those into happiness scores.[2] As Singer and MacAskill recommend the recommendations of GiveWell, a charity evaluator, I compare a mental health charity, StrongMinds, to GiveWell's current eight top-recommended charities, all of which target poverty or physical health. Mental health has not so far been considered as a potential priority either by GiveWell or Singer and

[1] See chapter 4. To be clear, I use 'making people happier' to refer to increasing someone's happiness, either by increasing the quality or length of their life. What happiness is discussed in chapter 4 also.
[2] See chapter 4 for an explanation and justification of this approach to measuring happiness.

MacAskill.[3] I show that StrongMinds looks competitive with the best of GiveWell's four life-improving charities (which focus on poverty). I run into problems comparing life-improving with life-saving charities due to the uncertainties about where to say the 'neutral point' equivalent to non-existence is on a 0—10 life satisfaction scale.

In so doing, this final chapter addresses some unresolved questions from chapter 4 and builds on the analysis of chapter 6.

In chapter 4, I raised four objections against using self-reported subjective well-being (SWB) scores to measure happiness, instead of the proxies that are currently favoured by moral philosophers (i.e. income and QALYs). While I tackled the first two objections there, I was unable to convincingly deal with the last two practical concerns: (3) there is not yet enough evidence on SWB to guide our decision-making; (4) moving to SWB would not make a practical difference and it is, therefore, unnecessary. The only way to fully address those concerns is to show that we can crunch the numbers and, when we do, it does make a difference to our priorities in at least one case. This is exactly what I do in this chapter. As mental health has not so far been thought of as a potential priority but looks competitive with the existing priority when we shine the light on self-reported happiness scores on matters, I think that is sufficient to address those two objections.

This chapter also picks up where the preceding chapter (6) ended. We had completed the 'cause mapping' survey of potentially high-impact ways to make people happier, but not gone to the next step of assessing if any of the as-yet-unexamined possibilities

---

[3] The notable exception is (Plant and Singer, 2017), a newspaper column, which argues (developed world) governments should urgently provide additional mental health treatment as doing so would alleviate suffering and reduce overall health expenditure. Singer informs me that StrongMinds will be mentioned in the second edition of his book *The Life You Can Save* (Singer, 2019) due out in Autumn (and this is due, in some part, to my agitation on the subject).

offered a most cost-effective way to do good than the existing priorities. Here, we take one step in that direction. GiveWell recommend developing world organisations that carry out evidence-based 'micro-interventions'—i.e. those that help people individually, as opposed to changing the system. As such, I choose to compare those against StrongMinds, another developing world organisation providing an evidence-based micro-intervention, because that allows the most straightforward 'apples-to-apples' analysis. I am not claiming interventions of this type are the most effective way to increase happiness, but it is easiest to start here.

I need to make a few preliminary comments before we get stuck into the cost-effectiveness analysis.

In a more ideal world, there would already have been randomised-controlled trials (RCTs) which had measured the impact of all the charities' interventions using measures of self-reported happiness. We could then look at that data and see, at a glance, if changing our measure of impact would change our priorities. Sadly, such information does not exist, nor is it likely to in the near future, and so I do the next best thing and create some estimates on the basis of currently available information.[4]

Specifically, my approach is to estimate the impact of different interventions in terms of life satisfaction point-years, abbreviated to 'LSPs', where 1 LSP is equivalent to increasing life satisfaction for one person by one point on a ten-point scale for one year. As noted earlier (chapter 4.2) some outcomes have a bigger impact on life satisfaction than on happiness (and vice versa). Hence, I then adjust the LSPs to get

[4] Among the economists using SWB measures I have spoken to, there appears to be almost no interest in using them to evaluate charity cost-effectiveness. If we want to wait for such information, we will be waiting for a long time.

happiness point-years, or 'HPs', where 1 HP is equivalent to increasing happiness for one person by one point on a ten-point scale for one year.

LSPs and HPs are similar in structure to QALYs, but with two differences.[5] First, LSPs and HPs are measured on a 0-10 scale, while QALYs are on a 0-1 scale (this is a trivial difference and I note it to reduce potential confusion). Second, while QALYs are measured on a ratio scale, life satisfaction and happiness are perhaps better understood having an interval scale. On the QALY, 0 is equivalent to having no health and thus also quite naturally equivalent to death. The typical scale that measures life satisfaction runs from 0 ("not at all satisfied") to 10 ("extremely satisfied"). Participants are not asked to specify what point on the 0 to 10 scale is the 'neutral point', where they are neither satisfied nor dissatisfied, which would be equivalent to non-existence, nor is a neutral point specified for them. The neutral point is arguably somewhere between 0 and 5 for life satisfaction, but I postpone discussion of this issue until section 3, when I set out the problems this uncertainty causes for assessing the cost-effectiveness of life-saving charities. As such, HPs and LSPs are not as straightforwardly interpretable as Well-being Adjust Life-Years (WALYs), which would, by specification, have a clear neutral point built-in.[6]

It will useful to briefly explain GiveWell's approach and how it differs from the SWB-based one I use. In short, GiveWell start with an estimate of how much it costs an organisation to achieve particular outcomes. The two most important outcomes in their analysis being (a) saving the life of an under-5-year old child and (b) doubling consumption for one person for one year. In order to trade-off those two outcomes

---

[5] See chapter 4.5 for a discussion of QALYs compared to SWB scores as a measure of happiness.
[6] As noted in chapter 4 at footnote 8, MacAskill suggests we would ideally measure the impact of our actions in WALYs.

against one another, they poll their staff, asking them how many years of doubled consumption are *morally equivalent* to saving the life of an under-5-year old child. GiveWell uses an aggregate of their staff's views to determine the trade-off between the outcomes—hence the 'GiveWell view' is a black box and no one's view in particular. In the recent analysis, the GiveWell view is that doubling someone's consumption for one year has 2% of the value of saving an under-5-year old.[7] To the best of my knowledge, no staff member makes their decisions based solely on SWB data, and only some take happiness to be the only intrinsic good. Hence the GiveWell approach is not (entirely) an attempt to maximise happiness. Nevertheless, for the purposes of my comparison, this is how I will interpret their view. I do not think this is grossly unfair, at least to an approximation: after all, Singer and MacAskill are looking for the most effective ways to increase happiness and recommend GiveWell's charities;[8] further, GiveWell's recommendations—which focus on saving lives and alleviating poverty—are intuitively highly plausible ways to increase happiness. Of course, if GiveWell were not at all interested in happiness—and instead, to illustrate the point with a silly example, were trying to maximise the total number of hats in the world—there would be little point in evaluating their recommendations in terms of their happiness impact. As it is, this seems a reasonable comparison to make.

I first compare StrongMinds against Givewell's life-improving recommendations and then against its life-saving recommendations.

---

[7] (GiveWell, 2018a)
[8] See chapter 4, footnotes 7 and 8.

2. StrongMinds vs GiveWell's recommended life-improving charities

I first estimate the cost-effectiveness of GiveDirectly, which provides unconditional cash transfers to poor villages in southern Africa, in terms of life satisfaction. I consider the impact of StrongMinds in terms of life satisfaction. I then convert their impact in terms of life satisfaction into their impact in terms of happiness. Finally, I attempt some comparisons between StrongMinds and GiveWell's other life-improving organisations and mention two further important considerations.

A study on GiveDirectly's cash transfers indicates they increase life satisfaction by about 0.3 LSPs after 4.3 months. I assume this effect lasts a whole year; it is the same for everyone in the recipient household, and there are 5 people per household on average. Hence, the annual LSP impact is 0.3 (LS/person) x 1 (year) x 5 (persons) = 1.5 LSPs. The average cash transfer is $750, implying a cost-effectiveness of 2 LSPs/$1000. The calculation and sources for this estimate are set out in table 7.1 at the end of the chapter.

We might wonder if there are impacts beyond one year. A 2018 study on the long-term (3-year) effects of GiveDirectly by Haushofer and Shapiro found that recipients, compared to non-recipients in distant villages, had 40% more assets but that recipients did no better on a psychological well-being index.[9] That the psychological well-being index didn't improve while assets did is admittedly surprising—research indicates wealth is associated with SWB.[10] Recalling the construct validation approach mentioned in section 3, I note one or more odd results shouldn't be taken as compelling evidence against the validity of SWB measures in general: someone making

[9] (Haushofer and Shapiro, 2018) Working paper. P. 22.
[10] E.g. see (Headey and Wooden, 2004)

such an objection would also need to explain how SWB measures could broadly get the 'right' answers whilst nevertheless failing to generally capture their underlying constructs.[11]

I note that my estimate of GiveDirectly's short-term impact on its recipients is large compared to what we would expect for that same relative increase in income to someone in the developed world. $1000 might be equivalent to a year's income for GiveDirectly's poorer recipients.[12] For someone in the UK, a doubling of income is associated with a 0.12 LSPs increase per year, whereas I infer GiveDirectly's per-person impact is 0.3, nearly three times larger. This doesn't seem that surprising, given the poverty of GiveDirectly's recipients. Extending this per-person effect across the whole family is likely generous to GiveDirectly: the association between family income and children's SWB has not been much examined, but one UK study found no statistically significant effects on children below 13 and only a small one for those above.[13]

Turning to StrongMinds, there have been no studies which directly measure its impact on life satisfaction, so I infer this using other available information. More detail is provided in table 7.1. The short explanation is that I use both UK data on life satisfaction and standardised mental health scores to estimate the life satisfaction impact of being treated for depression in the UK. I found this to be 0.3 LSPs in the first year. The treatment method used in the UK (cognitive-behavioural therapy) is different from the one used by StrongMinds (interpersonal therapy), but studies have

[11] I thank Peter Singer for identifying that this result could be taken as an objection as SWB measures in general.
[12] (Singer, 2015) at p. 113.
[13] (Knies, 2017)

found the methods are comparably effective.[14] However, as most of these studies were in a high-resource setting, I reduce StrongMinds' effectiveness by a third to be conservative.

The other important issue is how long the treatment lasts. A UK study of the long-term effects of psychotherapy found it lasted 4 years without substantial reduction—there was only 4 years of data.[15] Conservatively, this suggests that StrongMind's total effect is 0.8 LSPs per patient. In their analysis of StrongMinds' cost-effectiveness, Snowden et al. assumed the treatment effect had a 75% annual retention.[16] Modelled this way, the total effect is 0.75 LSPs per patient, which is very similar. I take the average of two estimates. StrongMinds say their per-participant costs are $102, which indicates their impact is 7.4 LSPs/$1000 (1 d.p.).

Hence, compared in terms of life satisfaction, StrongMinds is around 4 times more cost-effective: 2 LSPs/$1,000 vs 7.4 LSPs/$1,000.

In chapter 4.2, I noted some things have more effect on life satisfaction measures than happiness (and vice versa) and that once we had outcome measures in terms of life satisfaction, we could convert these into happiness scores by adjusting for these differences. The evidence indicates that mental health has about a 40% bigger impact on affect than life satisfaction,[17] and income increases have about a 60% smaller impact on affect than life satisfaction.[18] Hence, crudely, we could infer that StrongMinds' cost-effectiveness is 10.4 Happiness Point-Years (HPs)/$1,000 and

[14] (Cuijpers et al., 2013; Donker et al., 2013; Lemmens et al., 2015)
[15] (Wiles et al., 2016)
[16] (Halstead, Snowden and Heoijmakers, 2019)
[17] Inferred from (Dolan and Metcalfe, 2012): anxiety/depression level 3 has a 35% bigger impact on affect than life satisfaction; anxiety/depression level 2 has a 45% bigger impact on affect than life satisfaction. I take the average of these two numbers.
[18] (Boarini *et al.*, 2012) 24.

GiveDirectly's is 0.8 HPs/$1,000. If we make this adjustment, StrongMinds is then around 13 times more cost-effective than GiveDirectly in terms of *happiness*.

We can quickly compare StrongMinds to GiveWell's other life-improving charities—SCI, Deworm the World, SightSavers and END—all of which run deworming programmes where children are treated for intestinal worms.[19] On GiveWell's analysis, all these other charities are at least five times more cost-effective than GiveDirectly: Deworm the World is rated as 18.3 times better (the highest) and END 5.5 (the lowest).[20] Let's assume that GiveWell's analysis about the relative cost-effectiveness of these organisation is correct—there is not enough space to dispute that, and this is the charitable assumption. In that case, as StrongMinds is around 13 more cost-effective than GiveDirectly, it puts StrongMinds and Deworm the World *roughly* on a par in terms of happiness-based cost-effectiveness. Deworm the World is 40% more cost-effective on these numbers but, given the uncertainty that surrounds such estimates, we might think this is within the margin of error.

Before I move on, I want to note two further ways in which this analysis is likely to be generous to GiveWell's choices. First, it ignores the fact that making some people richer (and so happier and more satisfied) may have *negative spillovers*: it makes non-recipients feel poorer (and so less happy and satisfied). There seems to be mixed evidence that GiveDirectly has a negative spillover effect. One study found it was large enough to offset all the gains to life satisfaction.[21] However, GiveWell state that a more recent, not-yet-public study finds there are no negative SWB spillover effects across

[19] (GiveWell, no date b).
[20] (GiveWell, 2018a).
[21] (Haushofer, Reisinger and Shapiro, 2015).

villages and positive spillover effects to non-recipients within the same village.[22] The broad picture in the literature, as illustrated in figure 6.1, is that comparison effects do occur, hence it would be surprising if there were none at all. This concern will apply to GiveWell's other three life-improving organisations too: according to GiveWell, the vast majority of the benefit of deworming comes not from reducing the physical discomfort the worms cause, but from the fact that dewormed children do better in school and earn more in later life as a result.[23] Presumably, this extra income will also result in *some* negative spillover effects on the earners' peers.[24]

The second concern is one raised in chapters 1 and also relates to the potential downsides to making some richer: if people become richer, they eat more meat; assuming the animals suffer, and creating unhappy lives is bad, this will somewhat reduce the value of wealth-increasing interventions. Presumably, even if mental health interventions make people somewhat richer, e.g. because they can work more, poverty alleviation programmes would have a larger effect on income, which further relatively reduces the impact of the latter against the former.

I do not attempt to quantify these further considerations. Such matters are too complicated to resolve here. Even without doing so, I think the analysis is sufficient to show both that we can use happiness scores for cost-effectiveness and that doing so this indicates new potential priorities. It seems striking that by moving to a new outcome measure (self-reported happiness scores) we are able to find a type of

---

[22] See (GiveWell, no date a). GiveWell just state the effects in terms of 'subjective well-being' so I don't know what measure is being used.
[23] (GiveWell, 2018a) states that 2% of the cost-effectiveness of the deworming charities (DtW, SCI, Sightsavers, END) comes from 'short-term health effects' and 98% from 'eventual income and consumption gains'. See the Results tab in the spreadsheet.
[24] Peter Singer notes that improving education would also a further, societal benefit – better-educated workers boost the economy, increase employment, raise tax revenue, etc. This seems plausible and I merely concede I am unable to account for or quantify these potentially disperse benefits here.

intervention (treating mental health), one that had not been previously considered and is comparably cost-effective to the best of the existing interventions.

I anticipate a critical response here along the following lines: on closer inspection, my hopeful efforts to find a more cost-effective way to do good will turn out to be less promising than it currently seems, and no one should change where they donate their money until further evidence has been acquired.

I think this objection is probably correct. However, it is beside the point for the objections I am trying to counter, namely that the use of self-reported happiness scores to determine what increases happiness is either impractical or irrelevant. It seems the correct response to the concern I've raised is to more accurately assess charitable impact by using self-reported happiness scores and then see if the result will still hold.

3. StrongMinds vs GiveWell's recommended life-saving charities

The other important comparison to try to make is between StrongMinds and GiveWell's recommended life-saving interventions, such as the Against Malaria Foundation (AMF), which provides malaria-resistant bednets. For simplicity, this analysis ignores two major complexities discussed elsewhere in the thesis. First, that there are different philosophical views about how to assess the badness of death (see chapter 3). Second, the other-regarding impacts of saving lives (see chapters 1 and 2). Here, I use the *Deprivationism* account of the badness of death (see chapter 3), on which the value of saving a life is the total well-being that the person would have had if they'd lived.[25]

[25] If we wish to be more technical: the life comparative account of the badness of death could be restated as the additive account of lifetime well-being. That is the lifetime well-being value of a life is the sum of the momentary well-being which the individual has at each moment they exist.

How cost-effective is AMF? According to GiveWell's estimates, AMF saves a life (i.e. prevents a premature death) for around $4,500.[26] Suppose that grants 60 counterfactual years of life. To pick at random, average life satisfaction in Kenya, one country where AMF operates, is 4.4 out of 10.[27] Now we run into a problem. The typical scale that measures life satisfaction runs from 0 ("not at all satisfied") to 10 ("extremely satisfied"). But it's unclear where on the 0—10 is the 'neutral point' which we can take as being equivalent to non-existence. This is a deficiency of relying on life satisfaction scales for our measure of happiness: in contrast, measures of affect (i.e. happiness) usually specify a neutral point to represent the feeling of being, on balance, neither happy nor unhappy. Life satisfaction scales are of interval quality when we need a ratio scale to compare improving to saving lives.[28] To transform them to the latter we must assign a neutral point.

Prima face, the mid-point in the scale, 5, would be the neutral point. Yet, if that's true, then saving lives through AMF would, in fact, be bad: 4.4/10 is below the neutral point and thus, in saving lives, AMF would be prolonging lives not worth living.[29] It's not particularly plausible that those in Kenya have such bad lives.

Alternatively, as Clark et al. suggest, we could 'at a stretch' take 0 as being equivalent to having no life satisfaction.[30] One issue is that it then makes it impossible for someone to say they are dissatisfied with their life, even though such a view is intelligible. Clearly, it is possible to feel unhappy.[31] Here, we find an opposite problem:

---

[26] (GiveWell, 2019b).
[27] (Helliwell, Layard and Sachs, 2017) at p 28.
[28] See chapter 4.4 for discussion of different types of scales.
[29] This issue is complicated by the fact we might assume those lives will be happier and more satisfied over time. For simplicity, I leave out this issue.
[30] (Clark *et al.*, 2018) at p. 206.
[31] This may well be a limitation of using life satisfaction scores as a proxy for happiness.

if the neutral point is too low, it will erroneously count saving lives that are not worth living as good.

We could instead partially split the difference and say the neutral point is 4. If this is so, saving the child is worth 0.4 LSPs a year for 60 years, a total of 24 LSPs (0.4 x 60). Given the $3,500 cost, we can calculate cost-effectiveness as 6.9 LSPs/$1,000. Earlier, I estimated that StrongMinds' cost-effectiveness was 7.4 LSPs/$1000 and 11.8 HPs/$1,000. It's unclear what adjustment, if any, we should make for life-saving interventions when converting from life satisfaction to happiness, so I have made no adjustment and assumed that AMF's cost-effectiveness is 6.9 HPs/$1,000. If these estimates are correct, then StrongMinds is still more cost-effective, whether it is measured in terms of LSPs or HPs.

However, all this is sensitive to where the neutral point is placed: if the neutral point were 3/10, AMF's cost-effectiveness would leap to 24.4 LSPs/HPs per $1,000 and outperform StrongMinds. If the neutral point were 0, AMF would be about 75 HPs/LSPs per $1,000 and about seven times more cost-effective than StrongMinds. Note this analysis excludes any other-regarding impacts of saving lives.

If it were the case that life-improving interventions were clearly better or worse than life-saving ones on any (plausible) specification of the neutral point, this issue could be ignored. As it is, the comparison is highly sensitive to it. Further work is needed to determine the correct methodological approach to this issue. I am unsure what to suggest and am forced to leave the issue here.

4. Conclusion

In this chapter, I attempted a first-pass cost-effectiveness analysis of which charities are the best at making people happier.[32] While effective altruists have thought alleviating poverty is the best option, I argued that a mental health organisation, StrongMinds, seems to be roughly as cost-effective as the leading poverty-alleviating entities when we assess impact in terms of SWB. I was unable to satisfactorily compare this mental health intervention to life-saving physical health interventions.

Earlier (in chapter 4), I argued that SWB scores should be used to measure happiness but noted there were two practical objections to using SWB data—that judging outcomes in terms of self-reported happiness scores was both unfeasible and would not make a difference. As mental health interventions have not been previously identified as a potential happiness-increasing priority, the analysis in this chapter addresses those objections.

This chapter also made some progress by investigating one of the many possible ways of making people happier that was set out in chapter 6. There are many other options still to examine, but here, I was able to assess one novel option from that list and show it provides a means to do good *even* better.

[32] At least, people alive today.

Item	Number	Source	Note
GIVEDIRECTLY			
Increase in life satisfaction for recipients of GiveDirectly (standard deviations)	0.16	(Haushofer and Shapiro, 2016) p1976	Measured after 4.3 months. Note (Haushofer, Reisinger and Shapiro, 2015) at p32 show there is adaptation to the transfers—the initial effect reduces over time. I'm assuming the *average* effect over the year was a raise of 0.16 standard deviations. (I checked this with Julian Jamison, a development economist, who thought this was a reasonable approximation.)
Standard deviation of life satisfaction scores	1.9	(Clark *et al.*, 2018) p16	I could not find the standard deviation in (Haushofer and Shapiro, 2016) so use I UK data, assuming the standard deviation is the same—I have no reason to think it should be larger or smaller.
Life satisfaction increase per household member	0.304	Calculation	Multiplication of two previous numbers.
Members of household	5	Guess	2 parents, 3 children
Total life satisfaction effect on household (LS point-years)	1.52	Calculation	Assumes assumption all members of household had the measured effect size.
Cost-effectiveness of GiveDirectly in LS point-years/$1,000 (assuming no negative spillovers)	2.0	Calculation	To 1 d.p.
Cost-effectiveness of GiveDirectly in happiness point-years (HPs)	0.8	Calculation	(Boarini *et al.*, 2012) p24 find income increases have about a 60% smaller impact on affect than life satisfaction
STRONGMINDS			
Reduction in life satisfaction from suffering from depression/anxiety	0.7	(Clark *et al.*, 2018) p212.	LS impact of being diagnosed with mental illness vs not (from multiple regression)
Average PHQ-9 of those diagnosed with depression	15.5	(Kendrick *et al.*, 2009)	UK data.
Average PHQ-9 score of non-clinical population (i.e. those without depression)	5	(Gyani *et al.*, 2013)	Cut-off is 10. Assume the average is 5.

Difference in PHQ-9 scores between average person diagnosed with depression and non-clinical population	10.5	Calculation	
Average reduction in PHQ-9 score of treatment (UK)	4.47	(Gyani *et al.*, 2013)	From the UK's Improving Access to Psychological Therapies (IAPT) programme which provided CBT. Figure is from the treatment of moderate depression. This is conservative—those with more severe depression had larger average improvements.
StongMind own estimate of PHQ-9 reductions	4.5	(StrongMinds, 2015)	Note this was a quasi-RCT and conducted by StrongMinds. The 'control group' consisted of women who did not want the group therapy StrongMinds provides (rather than those who did want it).
Annual increase in life satisfaction from the treatment of depression/anxiety (UK)	0.298	Calculation	
First-year increase in life satisfaction caused by StrongMinds	0.2	Adjustment	Note StrongMinds's self-rated PHQ-9 reduction nearly identical to that seen in UK mental health treatment. I discount the effectiveness by 1/3 for conservatism.
Duration of the effect of mental health treatment (years)	4	(Wiles *et al.*, 2016)	Reduction in clinical scores for the UK's CBT appeared to be almost constant when measured 4 years later. Study only looked at a 4-year duration, so it is reasonable to assume the effect lasted longer. I assume 4 years for conservatism. An alternative would be to assume the effect fades as found in (Reay *et al.*, 2012)—see below.
Total LS effect, calculated as constant for 4 years	0.8	Calculation	
(Alternative calculation) Total LS effect, assuming there is a 75% annual retention of the benefit	0.75	Calculation	To check, I also calculated the total LS effect using the same method as (Halstead, Snowden and Heoijmakers, 2019), who assume a 75% annual retention of benefits based on the results of (Reay *et al.*, 2012). Numbers for this are below. The two methods give almost exactly the same results.

Per participant cost of StrongMinds ($)	102	(StrongMinds, 2018)	Total programme costs/no. of patients
Total LS effect of Strong (average of two duration calculations)	0.75	Calculation	Average of two estimates
Cost-effectiveness of StrongMinds in LS point-year/$1,000)	7.4	Calculation to 1 d. p.	
Cost-effectiveness of StrongMinds in happiness point-years (HPs)	10.4	Calculation to 1 d. p.	Inferred from (Dolan and Metcalfe, 2012): anxiety/depression level 3 has a 35% bigger impact on affect than life satisfaction; anxiety/depression level 2 has a 45% bigger impact on affect than life satisfaction. I take the average of these.
ALTERNATIVE RETENTION CALCULATION			
Retention rate of benefits	75%	(Reay *et al.*, 2012)	
Year 1 benefit (LS points)	0.2	Taken from the inferred 1st year of StrongMind's effect	
Year 2 benefits (LS points)	0.15	calculation	
Year 3 benefit (LS points)	0.1125	calculation	
Year 4 benefits (LS points)	0.0844	Calculation to 4 dp.	
Year 5 benefit (LS points)	0.0639	Calculation to 4 d.p.	
Year 6 benefits (LS points)	0.0475	Calculation to 4 d.p.	
Year 7 benefit (LS points)	0.0356	calculation to 4 d.p.	
Year 8 benefits (LS points)	0.0267	Calculation to 4 d.p.	
Year 9 benefits (LS points)	0.0200	Calculation to 4 d.p.	
Year 10 benefits (LS points)	0.0150	Calculation to 4 d.p.	
Total after 10 years (LS points)	0.7549	Calculation to 4 d.p.	

Table 7.1.

Conclusion

'What should we do if we want to do as much good as possible?' is a question that has only recently become the subject of serious inquiry. The effective altruism social movement has been responsible for a considerable portion of the analysis of it. In this thesis, I have argued that some of the claims by effective altruists are, to a greater or lesser extent, mistaken and shown how rectifying these errors could substantially alter both our understanding of, and our priorities for, doing the most good.

My *modus operandi* has been to take for granted the sort of basic assumptions philosophers normally argue about—assumptions about morality or facts—and then identify a whole range of considerations, given those assumptions, which have been overlooked so far and whose inclusion alters the analysis in interesting ways. As such, I have neither tried to argue what the correct theory of value is or argued what our priorities would be on such a theory, assuming that we wanted to do the most good. I hope to make such arguments in future work.

I'll now say, more specifically, what I think I have achieved, what I have not achieved and what I have not sought to do.

Chapter 1 drew out the underlying tension between two propositions: that saving humans is good and that being a meat eater is wrong because of the animal suffering this causes. I argued that, given certain other not implausible normative and empirical assumptions, saving strangers' lives is bad and not required. Although both Singer's essay *Famine, Affluence and Morality* (from which the famous *Drowning Child* case originates) and book *Animal Liberation* (from which the argument against meat eating arises) are decades old, the strong tension between them appears not to have

been brought out and evaluated before.[1] I did not attempt to argue that eating meat is wrong on the grounds that it creates animals with lives not worth living; rather, I assumed it for the sake of argument and then raised this problem for those that do.

Chapter 2 puts pressure on the acceptance of a different pair of beliefs. Many seem inclined towards the 'Intuitive View', that saving lives is good and that, due to concerns about overpopulation, averting lives is also good. I developed Greaves' earlier analysis of the topic. I showed how improbable the Intuitive View would be on Totalism and how, if the Intuitive View were true, neither saving nor averting lives would be particularly valuable. This raises a different challenge for Singer, who seems tempted to hold Totalism, the Intuitive View, and that saving and averting lives are charitable priorities for individuals who want to do the most good. I also explored this topic from a Person-Affecting View. A general challenge emerged: we don't know the value of saving or averting lives until we know whether and to what extent the Earth is under- or overpopulated. I was unable to say much about this question other than that where the Earth is in relation to its optimum population seems to be a highly complex, unclear empirical matter, regardless of whether one considers the effects on this generation or on all generations. I noted this uncertainty is relatively interesting, given how many people seem to assume it is either evident that the Earth is overpopulated or discussing this topic is morally unacceptable.

Chapter 3 took a different tack and considered whether, if we looked just at the self-regarding value of saving lives, saving lives—or, more specifically, saving children's lives in the developing world through highly cost-effective health interventions—could be the most good we could do. It seems obvious, at least to some, that this is the most

---

[1] (Singer, 1972), (Singer, 1975)

good we can do. I assessed this question from four different views of 'life-value' (a combination of a population axiology with a view of the badness of death) and presumed the majority of people would accept one of these views, or a variant of them. I argued that, for each view, either something else seemed more cost-effective than saving lives or it was not straightforward to make such an assessment due the indeterminacy of the view. As such, even if we set aside the considerations from chapters 1 and 2, many of those who are currently spending their resources saving lives should wonder if they can do more good elsewhere.

We might have believed it to be a settled issue, not just that saving lives was good, but that it was the most good you could do (at least, with your charitable donations). The upshot of the first three chapters is that matters are (frustratingly) more complex.

Having reached the limits of our analysis regarding saving lives, the attention of the thesis moved to considering how best to improve lives, that is, to make people happier during their lives. Chapter 4 is motivated by the realisation that, while social scientists have been measuring 'subjective well-being' (SWB) via self-reports for the last few decades, moral philosophers' suggestions about how to increase happiness have made little use of the findings produced by social scientists. In making their recommendations about which charities are most effective at increasing happiness, Singer and MacAskill, for instance, have tended to rely on other proxies for happiness, such as income and QALYs. This scenario is puzzling. I proposed four objections to using any type of self-reported SWB scores to measure happiness: (a) happiness can't be measured through self-reports; (b) happiness scores are not interpersonally cardinally comparable; (c) there isn't enough evidence of this type of guide decision making; and (d) using self-reports instead of current proxies would not change the priorities.

Having raised the objections, chapter 4 fully answers the first objection and partially answers the other three. Philosophers of science and social scientists have both previously argued happiness can be measured via self-reports. However, the arguments of these two disciplines are usually made in isolation. My contribution regarding the first objection is to stitch these arguments together into an answer that provides both an adequate theoretical and empirical justification. On the subject of interpersonal cardinality, I specify the six conditions are jointly sufficient for this (something that I do not believe has been done before). I then assess how reasonable it is to assume those conditions. I argue four of the six conditions do seem reasonably assumable, but there are doubts over the remaining two. I specified where further work would be needed and proposed that, in the meantime, as a practical matter, we should assume interpersonal cardinality. Finally, I argue that whether or not the 'raw' SWB scores are interpersonally cardinal is not, in itself, a sufficient reason not to use SWB scores—we can simply apply the relevant mathematical transformation so that the transformed scores become interpersonally cardinal. Hence, I hope to have provided a rigorous, somewhat reassuring, but necessarily incomplete, analysis on the use of SWB scores as a measure for happiness. Chapter 4 ends by noting that SWB scores indicate, at least *prima facie*, different priorities for increasing happiness than the proxies Singer and MacAskill relied on.

Chapter 5 asks, if we are going to reassess our priorities, what would be the best method to do this? Effective altruists, MacAskill observes, typically use a three-factor cause prioritisation framework for determining what the world's most pressing problems are. The framework seems not to have been thoroughly investigated. I take a closer look in order to try and address some open questions about its functioning; I suggest a moderate reconceptualisation. The main practical conclusion is that the

usefulness of the three-factor framework ultimately relies on how carefully we've thought about the particular solutions to the problems we're interested in.

In response to this conclusion, chapter 6 proposes a way of developing the extant cause prioritisation methodology to better find and organise promising solutions. I call this the 'cause mapping' approach. In essence, my suggestion is to break down the prioritisation process into smaller steps which can then be systematically worked through. I then apply cause mapping to the question of how a philanthropist could make people happier during their lives. I end up with a long list of potentially high-priorities options, many of which are novel (in the sense of not having been considered by effective altruists so far) and require further empirical investigation. As such, chapter 6 makes some progress towards creating an improved methodology for effective altruists and also highlight some new problems to examine.

Building on this work, in chapter 7 I evaluate the cost-effectiveness of a novel item on this long list—treating mental health. Drawing on the arguments from chapter 4 that happiness can be measured by self-reported SWB scores, I compare the cost-effectiveness of a developing world mental health charity, StrongMinds, against the charities recommended by GiveWell, a charity evaluator, using SWB scores. In my initial analysis, I find that StrongMinds seems roughly on a par with GiveWell's top life-improving charity and better than the rest. Although the analysis is somewhat simplistic and the numbers will presumably change on refinement, I nonetheless consider this result to be sufficient to meet the two practical objections to measuring happiness with SWB scores raised in chapter 4: that there isn't enough SWB-evidence to guide decision-making and using it would not give us new priorities. I think it is clear that moving to the measurement of happiness through self-reports opens promising new lines of inquiry in the pursuit of greater worldwide happiness.

The achievement of the thesis then is this. I have provided a sustained examination of three topics: the value of saving lives, how to best improve lives, and cause prioritisation methodology. In each case, I have brought new analysis and considerations to bear. These unsettle the apparent consensus around how to do the most good and, I hope, set out how we can do good even better.

Bibliography

80000 Hours (no date) *How to compare different global problems in terms of impact*. Available at: https://80000hours.org/articles/problem-framework/ (Accessed: 26 June 2017).

Adler, M. D., Dolan, P. and Kavetsos, G. (2017) 'Would you choose to be happy? Tradeoffs between happiness and the other dimensions of life in a large population survey', *Journal of Economic Behavior & Organization*, 139, pp. 60–73. doi: 10.1016/j.jebo.2017.05.006.

Ahuvia, A. (2008) 'If money doesn't make us happy, why do we act as if it does?', *Journal of Economic Psychology*, 29(4), pp. 491–507. doi: 10.1016/j.joep.2007.11.005.

Alderson, A. S. and Katz-Gerro, T. (2016) 'Compared to Whom? Inequality, Social Comparison, and Happiness in the United States', *Social Forces*. Aldine de Gruyter, New York, 95(1), pp. 25–54. doi: 10.1093/sf/sow042.

Alexandrova, A. (2017) *A Philosophy for the Science of Well-Being*. Oxford University Press. doi: 10.1093/oso/9780199300518.001.0001.

Alexandrova, A. and Haybron, D. M. (2016) 'Is Construct Validation Valid?', *Philosophy of Science*. University of Chicago PressChicago, IL, 83(5), pp. 1098–1109. doi: 10.1086/687941.

Angner, E. (2013a) 'Is Empirical Research Relevant to Philosophical Conclusions?', *Res Philosophica*, 90(3), pp. 365–385. doi: 10.11612/resphil.2013.90.3.4.

Angner, E. (2013b) 'Is it possible to measure happiness?: The argument from measurability', *European Journal for Philosophy of Science*, 3(2), pp. 221–240. doi:

10.1007/s13194-013-0065-2.

Animal Charity Evaluators (2016) *Why Farmed Animals?* Available at: https://animalcharityevaluators.org/donation-advice/why-farmed-animals/ (Accessed: 23 January 2019).

Animal Charity Evaluators (2018a) *Donation Impact*. Available at: https://animalcharityevaluators.org/donation-advice/donation-impact/ (Accessed: 11 August 2019).

Animal Charity Evaluators (2018b) *The Human League*. Available at: https://animalcharityevaluators.org/charity-review/the-humane-league/#overview (Accessed: 30 July 2019).

Anna Alexandrova (2016) 'Is well-being measurable after all?', *Public Health Ethics*. Narnia, 10(June), pp. 1–15. doi: 10.1111/1467-9973.00225.

Arrhenius, G. (2008) 'Life extension versus replacement', *Journal of Applied Philosophy*. Blackwell Publishing Ltd, 25(3), pp. 211–227. doi: 10.1111/j.1468-5930.2008.00413.x.

Arrhenius, G. (no date) *Population ethics: The challenge of future generations*.

Ayton, P., Pott, A. and Elwakili, N. (2007) 'Affective forecasting: Why can't people predict their emotions?', *Thinking & Reasoning*, 13(1), p. 62 <last_page> 80. doi: 10.1080/13546780600872726.

Bader, R. (forthcoming) 'Person-affecting utilitarianism', in *Oxford Handbook of Population Ethics*.

Batz, C. and Tay, Louis (2018) 'Gender Differences in Subjective Well-Being', in Diener, E., Oishi, S., and Tay, L (eds) *Handbook of Well-being*. Salt Lake: UT: DEF

Publishers, pp. 1–15. doi: nobascholar.com.

Beckstead, Nick (2013) *A Proposed Adjustment to the Astronomical Waste Argument*. Available at: https://www.lesswrong.com/posts/5czcpvqZ4RH7orcAa/a-proposed-adjustment-to-the-astronomical-waste-argument (Accessed: 23 January 2019).

Beckstead, Nicholas (2013) *On the overwhelming importance of shaping the far future*.

Bentham, J. (1789) *An introduction to the principles of morals and legislation*.

Bertram, C. and Rehdanz, K. (2015) 'The role of urban green space for human well-being', *Ecological Economics*, 120, pp. 139–152. doi: 10.1016/j.ecolecon.2015.10.013.

Bjørnskov, C. (2010) 'How comparable are the Gallup World Poll life satisfaction data?', *Journal of Happiness Studies*. Springer Netherlands, 11(1), pp. 41–60. doi: 10.1007/s10902-008-9121-6.

Blanchflower, D. G. and Oswald, A. J. (2004) 'Well-being over time in Britain and the USA', *Journal of Public Economics*, 88(7–8), pp. 1359–1386. doi: 10.1016/S0047-2727(02)00168-8.

Boarini, R. *et al.* (2012) *What Makes for a Better Life?*, *OECD Statistics Working Papers*.

Bond, T. and Lang, K. (2018) *The Sad Truth About Happiness Scales: Empirical Results*. Cambridge, MA. doi: 10.3386/w24853.

Boserup, E. (1981) *Population and technology*. Oxford: Blackwell.

Bostrom, N. (2003) 'Astronomical waste: The opportunity cost of delayed technological development', *Utilitas*, 15(03), pp. 308–314.

Bostrom, N. (2009) 'Pascal's mugging', *Analysis*, 69(3), pp. 443–445. doi: 10.1093/analys/anp062.

Bostrom, N. (2014) *Superintelligence*. OUP.

Bramble, B. (2018) *The passing of temporal well-being*, *The Passing of Temporal Well-Being*. doi: 10.4324/9781315213385.

Bronsteen, J., Buccafusco, C. J. and Masur, J. S. (2012) 'Well-Being Analysis vs. Cost-Benefit Analysis', *SSRN Electronic Journal*. doi: 10.2139/ssrn.1989202.

Broome, J. (2004) *Weighing Lives*. Oxford University Press. doi: 10.1093/019924376X.001.0001.

Bykvist, K. (2008) 'Violations of normative invariance: Some thoughts on shifty ought', *Theoria*. John Wiley & Sons, Ltd (10.1111), 73(2), pp. 98–120. doi: 10.1111/j.1755-2567.2007.tb01193.x.

Bykvist, K. (2017) 'Moral uncertainty', *Philosophy Compass*, 12(3), p. e12408. doi: 10.1111/phc3.12408.

Clark, A. E. *et al.* (2008) 'Lags and leads in life satisfaction: a test of the baseline hypothesis', *The Economic Journal*, 118(529), p. F243.

Clark, A. E. (2016) 'Adaptation and the Easterlin Paradox', in. Springer Japan, pp. 75–94. doi: 10.1007/978-4-431-55753-1_6.

Clark, A. E. (2017) 'Happiness, income and poverty', *International Review of Economics*. Springer Berlin Heidelberg, 64(2), pp. 145–158. doi: 10.1007/s12232-017-0274-7.

Clark, A. E. *et al.* (2018) *The origins of happiness : the science of well-being over the life course*.

Cohen, J. (1995) *How many people can the earth support?* New York; London: Norton.

Cotton-Barratt, O. (2016) *Prospecting for Gold*. Available at: https://www.effectivealtruism.org/articles/prospecting-for-gold-owen-cotton-barratt/ (Accessed: 17 April 2019).

Crisp, R. (2006) 'Hedonism reconsidered', *Philosophy and Phenomenological Research*, 73(3), pp. 619–645.

Crisp, R. and Chappell, T. (1998) 'Utilitarianism', *Routledge encyclopedia of philosophy*. 9th edn. London: Routledge.

Csikszentmihalyi, M. and Larson, R. (1987) 'Validity and reliability of the Experience-Sampling Method.', *The Journal of nervous and mental disease*, 175(9), pp. 526–536.

Cuijpers, P. *et al.* (2013) 'A Meta-Analysis of Cognitive-Behavioural Therapy for Adult Depression, Alone and in Comparison with other Treatments', *The Canadian Journal of Psychiatry*, 58(7), pp. 376–385. doi: 10.1177/070674371305800702.

Cukrowicz, K. and Joiner, T. (2007) 'Computer-based intervention for anxious and depressive symptoms in a non-clinical population', *Cognitive Therapy and Research*.

Deaton, A. (2012) 'The financial crisis and the well-being of Americans', *Oxford Economic Papers*. Cambridge, MA, 64(1), pp. 1–26. doi: 10.1093/oep/gpr051.

Deaton, A. and Stone, A. A. (2013) 'Two Happiness Puzzles', *American Economic Review*, 103(3), pp. 591–597. doi: 10.1257/aer.103.3.591.

Van De Deijl, W. (2017) 'Which Problem of Adaptation?', *Utilitas*. Cambridge University Press, 29(4), pp. 474–492. doi: 10.1017/S0953820816000431.

Delgado, C. L. (2003) 'Rising Consumption of Meat and Milk in Developing Countries

Has Created a New Food Revolution', *The Journal of Nutrition*. Oxford University Press, 133(11), pp. 3907S-3910S. doi: 10.1093/jn/133.11.3907S.

Dhondt, A. (1988) 'Carrying capacity: a confusing concept', *Acta Oecologica* , 9(4), pp. 337–346.

Dickens, M. (2016) *Evaluation Frameworks (or: When Importance / Neglectedness / Tractability Doesn't Apply)*. Available at: http://mdickens.me/2016/06/10/evaluation_frameworks_(or-_when_scale-neglectedness-tractability_doesn't_apply)/ (Accessed: 14 January 2018).

Diener, E. *et al.* (1999) 'Subjective well-being: Three decades of progress', *Psychological Bulletin*, pp. 276–302. doi: 10.1037/0033-2909.125.2.276.

Diener, E., Kahneman, D., *et al.* (2010) 'Income's Association with Judgments of Life Versus Feelings', in *International Differences in Well-Being*. Oxford University Press, pp. 3–15. doi: 10.1093/acprof:oso/9780199732739.003.0001.

Diener, E., Wirtz, D., *et al.* (2010) 'New Well-being Measures: Short Scales to Assess Flourishing and Positive and Negative Feelings', *Social Indicators Research*. Springer Netherlands, 97(2), pp. 143–156. doi: 10.1007/s11205-009-9493-y.

Diener, E., Lucas, R., *et al.* (2010) *Well-Being for Public Policy*, *Well-Being for Public Policy*. Oxford University Press. doi: 10.1093/acprof:oso/9780195334074.001.0001.

Diener, E. *et al.* (2017) 'Findings all psychologists should know from the new science on subjective well-being.', *Canadian Psychology/Psychologie canadienne*, 58(2), pp. 87–104. doi: 10.1037/cap0000063.

Diener, E., Inglehart, R. and Tay, L. (2013) 'Theory and Validity of Life Satisfaction Scales', *Social Indicators Research*. Springer Netherlands, 112(3), pp. 497–527. doi:

10.1007/s11205-012-0076-y.

Diener, E., Lucas, R. E. and Oishi, S. (2018) 'Advances and Open Questions in the Science of Subjective Well-Being', *Collabra: Psychology*, 4(1), p. 15. doi: 10.1525/collabra.115.

Diener, E., Lucas, R. E. and Scollon, C. N. (2009) 'Beyond the hedonic treadmill: Revising the adaptation theory of well-being', in. Springer (The science of well-being), pp. 103–118.

Doepke, M. (2005) 'Child mortality and fertility decline: Does the Barro-Becker model fit the facts?', *Journal of Population Economics*. Springer-Verlag, 18(2), pp. 337–366. doi: 10.1007/s00148-004-0208-z.

Dolan, P. and Metcalfe, R. (2012) 'Valuing health: a brief report on subjective well-being versus preferences', *Medical decision making : an international journal of the Society for Medical Decision Making*, 32(4), pp. 578–582. doi: 10.1177/0272989X11435173 [doi].

Dolan, P., Peasgood, T. and White, M. (2008) 'Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being', *Journal of Economic Psychology*, 29(1), pp. 94–122. doi: 10.1016/j.joep.2007.09.001.

Dolan, P. and White, M. P. (2007) 'How Can Measures of Subjective Well-Being Be Used to Inform Public Policy?', *Perspectives on Psychological Science*. SAGE PublicationsSage CA: Los Angeles, CA, 2(1), pp. 71–85. doi: 10.1111/j.1745-6916.2007.00030.x.

Donker, T. *et al.* (2013) 'Internet-delivered interpersonal psychotherapy versus internet-delivered cognitive behavioral therapy for adults with depressive symptoms:

Randomized controlled noninferiority trial', *Journal of Medical Internet Research*, 15(5), p. e82. doi: 10.2196/jmir.2307.

Easterlin, R. A. (2016) 'Paradox Lost?', *SSRN Electronic Journal*. doi: 10.2139/ssrn.2714062.

Edgeworth, F. Y. (1881) *Mathematical Psychics*. London: Kegan Paul.

Edwards, E. (1964) 'On the theory of scales of measurement', *Ergonomics*. American Association for the Advancement of Science, pp. 504–505. doi: 10.1080/00140136408956259.

Ehde, D. M., Dillworth, T. M. and Turner, J. A. (2014) 'Cognitive-behavioral therapy for individuals with chronic pain: Efficacy, innovations, and directions for research', *American Psychologist*, 69(2), pp. 153–166. doi: 10.1037/a0035747.

Ehrlich, P. (1978) *The population bomb*. Ballant Books.

Ekers, D. *et al.* (2014) 'Behavioural activation for depression; An update of meta-analysis of effectiveness and sub group analysis', *PLoS ONE*. Edited by A. Aleman. Public Library of Science, 9(6), p. e100100. doi: 10.1371/journal.pone.0100100.

Epicurus (2019) *Letter to Menoeceus*. Translated by Hicks, R. Available at: http://classics.mit.edu/Epicurus/menoec.html (Accessed: 23 January 2019).

Etilé, F. *et al.* (2017) 'Modelling Heterogeneity in the Resilience to Major Socioeconomic Life Events', *PSE Working Papers*. HAL.

Fehige, C. and Wessels, U. (1998) 'Preferences: An introduction', in Wessels, U. and Fehige, C. (eds) *Preferences*. Berlin: W. de Gruyter, pp. xx–xliii.

Feldman, F. (2010a) 'On the philosophical implications of empirical research on happiness', *Social Research*. The Johns Hopkins University Press, 77(2), pp. 625–658.

doi: 10.2307/40972232.

Feldman, F. (2010b) *What is This Thing Called Happiness?* Oxford University Press.

Ferrer-i-Carbonell, A. and Frijters, P. (2004) 'How Important is Methodology for the estimates of the determinants of Happiness?*', *The Economic Journal*. Blackwell Publishing Ltd, 114(497), pp. 641–659. doi: 10.1111/j.1468-0297.2004.00235.x.

Frederick, S. and Loewenstein, G. (1999) 'Hedonic Adaptation', in *Well-being: The foundations of hedonic psychology*, pp. 302–329.

Frick, J. (2017) 'On the survival of humanity', *Canadian Journal of Philosophy*. Routledge, 47(2–3), pp. 344–367. doi: 10.1080/00455091.2017.1301764.

Frowe, H. (2018) 'Lesser-Evil Justifications for Harming: Why We're Required to Turn the Trolley', *The Philosophical Quarterly*. Oxford University Press, 68(272), pp. 460–480. doi: 10.1093/pq/pqx065.

Gamlund, E. and Solberg, C. T. (2019) *Saving people from the harm of death*. OUP.

George, M. S. *et al.* (1995) 'Daily repetitive transcranial magnetic stimulation (rTMS) improves mood in depression.', *Neuroreport*, 6(14), pp. 1853–6.

Gilbert, D. T. *et al.* (1998) 'Immune Neglect: A Source of Durability Bias in Affective Forecasting', *Journal of Personality and Social Psychology*, 75(3), pp. 617–638. doi: 10.1037/0022-3514.75.3.617.

GiveWell (2014) *David Roodman's draft writeup on the mortality-fertility connection - The GiveWell Blog*. Available at: http://blog.givewell.org/2014/04/17/david-roodmans-draft-writeup-on-the-mortality-fertility-connection/ (Accessed: 21 April 2017).

GiveWell (2018a) *2018 GiveWell Cost-Effectiveness Analysis — Version 4*. Available

at: https://docs.google.com/spreadsheets/d/1moyxmsn4UjhH3CzFJmPwAN7LUAkmmaoXDb6bdW3WILg/edit#gid=1364064522.

GiveWell (2018b) *GiveWell's Impact*. Available at: https://www.givewell.org/about/impact (Accessed: 2 August 2018).

GiveWell (2019a) *2019 GiveWell Cost-effectiveness Analysis — Version 1 - Google Sheets*. Available at: https://docs.google.com/spreadsheets/d/1k-lBNYFDB43EXVFnSozRBnNQnenjBwdmzaXKOjoFJ0M/edit#gid=1364064522 (Accessed: 23 January 2019).

GiveWell (2019b) *2019 GiveWell Cost-effectiveness Analysis — Version 5*. Available at: https://docs.google.com/spreadsheets/d/1d255LKz11L3V-OgOEns9WvJzpnVeaLTcEP1HD4lC478/edit#gid=1364064522 (Accessed: 11 August 2019).

GiveWell (no date a) *Spillover Effects of GiveDirectly's Cash Transfers Program | GiveWell*. Available at: https://www.givewell.org/international/technical/programs/cash-transfers/spillovers (Accessed: 3 June 2019).

GiveWell (no date b) *Top Charities*. Available at: https://www.givewell.org/charities/top-charities (Accessed: 23 January 2019).

GlobalData (2017) *Top Trends in Prepared Foods 2017: Exploring trends in meat, fish and seafood; pasta, noodles and rice; prepared meals; savory deli food; soup; and meat substitutes*.

Gold, M. R., Stevenson, D. and Fryback, D. G. (2002) 'HALYs and QALYs and DALYs, Oh My: Similarities and Differences in Summary Measures of Population Health',

Annual Review of Public Health. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA, 23(1), pp. 115–134. doi: 10.1146/annurev.publhealth.23.100901.140513.

Graham, L. and Oswald, A. J. (2010) 'Hedonic capital, adaptation and resilience', *Journal of Economic Behavior and Organization*. North-Holland, 76(2), pp. 372–384. doi: 10.1016/j.jebo.2010.07.003.

Greaves, H. (2015) 'The Social Disvalue of Premature Deaths', in *Weighing and Reasoning*. Oxford University Press, pp. 72–86. doi: 10.1093/acprof:oso/9780199684908.003.0006.

Greaves, H. (2017) 'Population Axiology', *Philosophy Compass*.

Greaves, H. (2019) 'Against "the badness of death"', in Gamlund, E. and Solberg, C. T. (eds) *Saving Lives from the Badness of Death*. Oxford University Press.

Greaves, H. (no date) 'Optimum population size', in Arrhenius, Bykvist, and Campbell (eds) *Oxford handbook of population ethics*. Oxford University Press.

Greaves, H. and Ord, T. (2017) 'Moral Uncertainty About Population Axiology', *Journal of Ethics and Social Philosophy*, 12(2), pp. 135–167. doi: 10.26556/jesp.v12i2.223.

Gyani, A. *et al.* (2013) 'Enhancing recovery rates: lessons from year one of IAPT.', *Behaviour research and therapy*. Elsevier, 51(9), pp. 597–606. doi: 10.1016/j.brat.2013.06.004.

Halstead, J., Snowden, J. and Heoijmakers, S. (2019) *Cause Report - Mental Health*.

Harmon-Jones, E., Gable, P. A. and Price, T. F. (2013) 'Does Negative Affect Always Narrow and Positive Affect Always Broaden the Mind? Considering the Influence of

Motivational Intensity on Cognitive Scope', *Current Directions in Psychological Science*. SAGE PublicationsSage CA: Los Angeles, CA, 22(4), pp. 301–307. doi: 10.1177/0963721413481353.

Haushofer, J., Reisinger, J. and Shapiro, J. (2015) 'Your Gain Is My Pain: Negative Psychological Externalities of Cash Transfers, Working Paper'.

Haushofer, J. and Shapiro, J. (2016) 'The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya', *The Quarterly Journal of Economics*. Oxford University Press, 131(4), pp. 1973–2042. doi: 10.1093/qje/qjw025.

Haushofer, J. and Shapiro, J. (2018) *The long-term impact of unconditional case transfers: expiremental evidence from Kenya*.

Haybron, D. M. (2016) 'Mental State Approaches to Well-Being', in Adler, M. D. and Fleurbaey, M. (eds) *The Oxford Handbook of Well-Being and Public Policy*. Oxford University Press. doi: 10.1093/oxfordhb/9780199325818.013.11.

Headey, B. and Wooden, M. (2004) 'The effects of wealth and income on subjective well-being and ill-being', *Economic Record*. John Wiley & Sons, Ltd (10.1111), 80(SPEC. ISS.), pp. S24–S33. doi: 10.1111/j.1475-4932.2004.00181.x.

Helliwell, J. *et al.* (2009) *International evidence on the social context of well-being*. w14720.

Helliwell, J., Bonikowska, A. and Shiplett, H. (2016) *Migration as a Test of the Happiness Set Point Hypothesis: Evidence from Immigration to Canada*. Cambridge, MA. doi: 10.3386/w22601.

Helliwell, J. F., Layard, R. and Sachs, J. (2017) *World happiness report 2017*. Sustainable Development Solutions Network.

Helliwell, J., Layard, R. and Sachs, J. (2018) *World Happiness Report 2018*.

Helliwell, J., Layard, R. and Sachs, J. (2019) *World Happiness Report 2019*.

Heyd, D. (2009) 'The intractability of the nonidentity problem', in. Springer (Harming future persons), pp. 3–25.

Heyd, D. (2014) 'Parfit on the Non-identity Problem, Again', *The Law & Ethics of Human Rights*. De Gruyter, 8(1), pp. 1–20. doi: 10.1515/lehr-2014-0003.

J, E. (1991) 'The meat factory: cruel cost of cheap pork and poultry – factory methods have slashed meat prices in the last 30 years', *The Guardian*, 14 October.

Jebb, A. T. *et al.* (2018) 'Happiness, income satiation and turning points around the world', *Nature Human Behaviour*, 2(1). doi: 10.1038/s41562-017-0277-0.

Jesteadt, W., Wier, C. C. and Green, D. M. (1977) 'Intensity discrimination as a function of frequency and sensation level', *The Journal of the Acoustical Society of America*. Acoustical Society of America, 61(1), pp. 169–177. doi: 10.1121/1.381278.

Kagan, S. (2011) 'Do I Make a Difference?', *Philosophy and Public Affairs*. Wiley/Blackwell (10.1111), 39(2), pp. 105–141. doi: 10.1111/j.1088-4963.2011.01203.x.

Kagan, S. (2016) 'What's Wrong with Speciesism? (Society of Applied Philosophy Annual Lecture 2015)', *Journal of Applied Philosophy*, 33(1), pp. 1–21. doi: 10.1111/japp.12164.

Kahneman, D. *et al.* (1993) 'When More Pain Is Preferred to Less: Adding a Better End', *Psychological Science*. SAGE PublicationsSage CA: Los Angeles, CA, 4(6), pp. 401–405. doi: 10.1111/j.1467-9280.1993.tb00589.x.

Kahneman, D. *et al.* (2004) 'A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method', *Science*, 306(5702), pp. 1776–1780.

doi: 10.1126/science.1103572.

Kahneman, D. *et al.* (2006) 'Would you be happier if you were richer? A focusing illusion', *Science*, 312(5782), pp. 1908–1910.

Kahneman, D. (2011) *Thinking, fast and slow*. Macmillan.

Kahneman, D. and Deaton, A. (2010) 'High income improves evaluation of life but not emotional well-being', *Proceedings of the National Academy of Sciences of the United States of America*, 107(38), pp. 16489–16493. doi: 10.1073/pnas.1011492107 [doi].

Kahneman, D. and Krueger, A. B. (2006) 'Developments in the Measurement of Subjective Well-Being', *Journal of Economic Perspectives*, 20(1), pp. 3–24. doi: 10.1257/089533006776526030.

Kamm, F. (1998) *Morality, Mortality: Volume I*. Oxford Univ. Press.

Kamm, F. M. (2001) *Morality, Mortality Volume II: Rights, Duties, and Status*. Oxford University Press. doi: 10.1093/0195144023.001.0001.

Kamm, F. M. (2019) 'The Badness of Death and What to Do about It (if Anything)', in *Saving People from the Harm of Death*. Oxford University Press, pp. 146–162. doi: 10.1093/oso/9780190921415.003.0011.

Kant, I. (1798) *Anthropology, History, and Education*. Edited by R. Louden and G. Zoller. Cambridge: Cambridge University Press. doi: doi:10.1017/CBO9780511791925.

Kaviani, H. (2011) 'Mindfulness-based cognitive therapy (MBCT) reduces depression and anxiety induced by real stressful setting in non-clinical population', *Revista Internacional de Psicología y Terapia*.

Kendrick, T. *et al.* (2009) 'Management of depression in UK general practice in relation to scores on depression severity questionnaires: analysis of medical record

data.', *BMJ (Clinical research ed.)*. British Medical Journal Publishing Group, 338, p. b750. doi: 10.1136/BMJ.B750.

Kennedy, S. H. *et al.* (2011) 'Deep Brain Stimulation for Treatment-Resistant Depression: Follow-Up After 3 to 6 Years', *American Journal of Psychiatry*. American Psychiatric Publishing Arlington, VA, 168(5), pp. 502–510. doi: 10.1176/appi.ajp.2010.10081187.

Keynes, J. M. (1923) *A tract on monetary reform*.

Khoury, B. *et al.* (2013) 'Mindfulness-based therapy: a comprehensive meta-analysis', *Clinical psychology review*, 33(6), pp. 763–771. doi: 10.1016/j.cpr.2013.05.005 [doi].

Kim, H. *et al.* (2017) 'Social comparison, personal relative deprivation, and materialism', *British Journal of Social Psychology*, 56(2), pp. 373–392. doi: 10.1111/bjso.12176.

Knies, G. (2017) 'Income effects on children's life satisfaction: Longitudinal Evidence for England'.

Koivumaa-Honkanen, H. *et al.* (2001) 'Life Satisfaction and Suicide: A 20-Year Follow-Up Study', *American Journal of Psychiatry*, 158(3), pp. 433–439. doi: 10.1176/appi.ajp.158.3.433.

Kristoffersen, I. (2011) *The Subjective Wellbeing Scale: How Reasonable is the Cardinality Assumption?*, *Economics Discussion / Working Papers*. The University of Western Australia, Department of Economics.

Kristoffersen, I. (2017) 'The Metrics of Subjective Wellbeing Data: An Empirical Evaluation of the Ordinal and Cardinal Comparability of Life Satisfaction Scores', *Social Indicators Research*. Springer Netherlands, 130(2), pp. 845–865. doi:

10.1007/s11205-015-1200-6.

Krueger, A. B. and Schkade, D. A. (2008) 'The reliability of subjective well-being measures', *Journal of Public Economics*. Elsevier, 92(8–9), pp. 1833–1845. doi: 10.1016/j.jpubeco.2007.12.015.

Lau, A. (2007) 'Measurement of Subjective Wellbeing: Cultural Issues', in *Proceedings of the 9th Australian Conference on Quality of Life*. Deakin University.

Layard, R. (2003) 'Happiness: has social science a clue? Lecture 1: what is happiness? Are we getting happier?', in *Lionel Robbins memorial lecture series*.

Lazari-Radek, K. de and Singer, P. (2014) *The Point of View of the Universe*. Oxford University Press. doi: 10.1093/acprof:oso/9780199603695.001.0001.

Lemmens, L. H. J. M. *et al.* (2015) 'Clinical effectiveness of cognitive therapy v. interpersonal psychotherapy for depression: Results of a randomized controlled trial', *Psychological Medicine*, 45(10), pp. 2095–2110. doi: 10.1017/S0033291715000033.

Lewis, D. (1986) *On the plurality of worlds*. B. Blackwell.

Lewis, G. (2018) *The person-affecting value of existential risk reduction*, *Effective Altruism Forum*. Available at: http://effective-altruism.com/ea/1n0/the_personaffecting_value_of_existential_risk/ (Accessed: 21 September 2018).

Lischetzke, T. and Eid, M. (2006) 'Why extraverts are happier than introverts: The role of mood regulation', *Journal of Personality*, 74(4), pp. 1127–1162. doi: 10.1111/j.1467-6494.2006.00405.x.

Luhmann, M. *et al.* (2012) 'Subjective well-being and adaptation to life events: a meta-analysis.', *Journal of personality and social psychology*, 102(3), p. 592.

MacAskill, W. (2015) *Doing Good Better*. Faber & Faber.

MacAskill, W. (2018) 'Understanding Effective Altruism and Its Challenges', in *The Palgrave Handbook of Philosophy and Public Policy*. Cham: Springer International Publishing, pp. 441–453. doi: 10.1007/978-3-319-93907-0_34.

Malthus, T. (1798) 'An essay on the principle of population'.

Maslen, H. *et al.* (2013) 'Regulation of devices for cognitive enhancement', *The Lancet*. Elsevier, pp. 938–939. doi: 10.1016/S0140-6736(13)61931-5.

Matheny, G. and Chan, K. M. A. (2005) 'Human Diets and Animal Welfare: the Illogic of the Larder', *Journal of Agricultural and Environmental Ethics*. Kluwer Academic Publishers, 18(6), pp. 579–594. doi: 10.1007/s10806-005-1805-x.

McMahan, J. (2002) *The ethics of killing: Problems at the margins of life*.

McMahan, J. (2008) 'Eating animals the nice way', *Daedalus*. MIT Press 238 Main St., Suite 500, Cambridge, MA 02142-1046 USA journals-info@mit.edu, 137(1), pp. 66–75. doi: 10.1162/daed.2008.137.1.66.

McMahan, J. (2009) 'Asymmetries in the Morality of Causing People to Exist', in David Wasserman and Melinda Roberts (eds) *Harming Future Persons*, pp. 49–68. doi: 10.1007/978-1-4020-5697-0_3.

McMahan, J. (2019) 'Early Death and Later Suffering', in *Saving People from the Harm of Death*. Oxford University Press, pp. 116–133. doi: 10.1093/oso/9780190921415.003.0009.

Melinda Gates (2014) *Annual Letter 2014 - Bill & Melinda Gates Foundation*. Available at: http://www.gatesfoundation.org/Who-We-Are/Resources-and-Media/Annual-Letters-List/Annual-Letter-2014 (Accessed: 21 April 2017).

Mercy for Animals (2011) *Farm to Fridge - The Truth Behind Meat Production*. Available at: https://www.youtube.com/watch?v=THIODWTqx5E (Accessed: 18 June 2018).

Michalos, A. C. and Maurine Kahlke, P. (2010) 'Stability and Sensitivity in Perceived Quality of Life Measures: Some Panel Results', *Social Indicators Research*, 98(3), pp. 403–434. doi: 10.1007/s11205-009-9554-2.

Mill, J. (1861) *Utilitarianism*.

Morewedge, C. K. and Buechel, E. C. (2013) 'Motivated underpinnings of the impact bias in affective forecasts', *Emotion (Washington, D.C.)*, 13(6), pp. 1023–1029. doi: 10.1037/a0033797 [doi].

Nagel, T. (1995) *Equality and Partiality*. Oxford University Press. doi: 10.1093/0195098390.001.0001.

Narveson, J. (1973) 'Moral problems of population', *The Monist*, pp. 62–86.

National Chicken Council (2018) *Chickopedia: What Consumers Need to Know - The National Chicken Council*. Available at: https://www.nationalchickencouncil.org/about-the-industry/chickopedia/#two (Accessed: 18 June 2018).

De Neve, J. E. *et al.* (2012) 'Genes, economics, and happiness', *Journal of Neuroscience, Psychology, and Economics*. NIH Public Access, 5(4), pp. 193–211. doi: 10.1037/a0030292.

Ng, Y.-K. (1995) 'Towards welfare biology: Evolutionary economics of animal consciousness and suffering', *Biology and Philosophy*, 10(3), pp. 255–285. doi: 10.1007/BF00852469.

Ng, Y. (1997) 'A case for happiness, cardinalism, and interpersonal comparability', *The Economic Journal*, 107(445), pp. 1848–1858.

Ng, Y. K. (2008) 'Happiness studies: Ways to improve comparability and some public policy implications', *Economic Record*. John Wiley & Sons, Ltd (10.1111), 84(265), pp. 253–266. doi: 10.1111/j.1475-4932.2008.00466.x.

Nichols, D. E., Johnson, M. W. and Nichols, C. D. (2017) 'Psychedelics as Medicines: An Emerging New Paradigm', *Clinical Pharmacology and Therapeutics*, 101(2). doi: 10.1002/cpt.557.

Norheim, O. (2019) 'The Badness of Death', in *Saving People from the Harm of Death*. Oxford University Press, pp. 33–47. doi: 10.1093/oso/9780190921415.003.0003.

Nozick, R. (1974) *Anarchy, state, and utopia*. New York: Basic Books.

Nussbaum, M. C. (2012) 'Who is the happy warrior? Philosophy, happiness research, and public policy', *International Review of Economics*. Springer-Verlag, 59(4), pp. 335–361. doi: 10.1007/s12232-012-0168-7.

Nutt, D. J., King, L. A. and Nichols, D. E. (2013) 'Effects of Schedule I drug laws on neuroscience research and treatment innovation', *Nature Reviews Neuroscience*, 14(8), pp. 577–585. doi: 10.1038/nrn3530.

OECD (2013) *Guidelines on Measuring Subjective Well-being*. OECD Publishing. doi: 10.1787/9789264191655-en.

ONS (2011) *Initial investigation into Subjective Wellbeing from the Opinions Survey*.

ONS (2018) *Personal well-being in the UK*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/measuringnationalwellbeing/july2017tojune2018 (Accessed: 28 May 2019).

Open Philanthropy Project and Karnofsky, H. (2014) *Narrowing down U.S. policy areas*. Available at: https://blog.givewell.org/2014/05/22/narrowing-down-u-s-policy-areas/ (Accessed: 14 January 2018).

Oswald, A. J. and Powdthavee, N. (2008) 'Death, happiness, and the calculation of compensatory damages', *The Journal of Legal Studies*, 37(S2), p. S251.

Page, R. (2001) 'TreeView', *Glasgow University, Glasgow, UK*.

Parfit, D. (1984) *Reasons and persons*. OUP Oxford.

Parfit, D. (1986) 'Overpopulation and the Quality of Life', in Singer, P. (ed.) *Applied Ethics*. Oxford University Press.

Parfit, D. (2011) *On what matters Volume one*. Oxford University Press.

Pavot, W. (2018) 'The Cornerstone of Research on Subjective Well-Being: Valid Assessment Methodology', *Handbook of Well-Being*. Edited by E. Diener, S. Oishi, and L. Tay. Salt Lake City: UT: DEF Publishers, pp. 83–93.

Perez-Truglia, R. (2012) 'On the causes and consequences of hedonic adaptation', *Journal of Economic Psychology*, 33(6), pp. 1182–1192. doi: 10.1016/j.joep.2012.08.004.

Plant, M. (2019) *Research agenda - Happier Lives Institute*. Available at: https://www.happierlivesinstitute.org/research-agenda.html (Accessed: 4 July 2019).

Plant, M. and Singer, P. (2017) 'The Moral Urgency of Mental Health', *Project Syndicate*, 16 November.

van Praag, B. M. S. (1991) 'Ordinal and cardinal utility. An integration of the two dimensions of the welfare concept', *Journal of Econometrics*. North-Holland, 50(1–

2), pp. 69–89. doi: 10.1016/0304-4076(91)90090-Z.

van Praag, B. M. S. (1993) 'The Relativity of the Welfare Concept', in *The Quality of Life*. Oxford University Press, pp. 362–385. doi: 10.1093/0198287976.003.0027.

Pummer, T. and MacAskill, W. (2019) 'Effective Altruism', *International Encyclopedia of Ethics*.

Rayo, L. and Becker, G. S. (2007) 'Evolutionary Efficiency and Happiness', *Journal of Political Economy*, 115(2), pp. 302–337. doi: 10.1086/516737.

Reay, R. E. *et al.* (2012) 'Trajectories of long-term outcomes for postnatally depressed mothers treated with group interpersonal psychotherapy', *Archives of Women's Mental Health*. Springer Vienna, 15(3), pp. 217–228. doi: 10.1007/s00737-012-0280-4.

Ren, J. *et al.* (2014) 'Repetitive transcranial magnetic stimulation versus electroconvulsive therapy for major depression: A systematic review and meta-analysis', *Progress in Neuro-Psychopharmacology and Biological Psychiatry*. Elsevier, 51, pp. 181–189. doi: 10.1016/J.PNPBP.2014.02.004.

Robbins, L. (1932) *An essay on the nature and significance of economic science,*. London: Macmillan.

Sachs, J. *et al.* (2019) *Global Happiness and Wellbeing Policy Report 2019*.

Savulescu, J. and Kahane, G. (2009) 'The moral obligation to create children with the best chance in life', *Bioethics*. Blackwell Publishing Ltd, 23(5), pp. 274–290. doi: 10.1111/j.1467-8519.2008.00687.x.

Schwarz, N. (1995) 'What Respondents Learn from Questionnaires: The Survey Interview and the Logic of Conversation', *International Statistical Review / Revue*

Internationale de Statistique, 63(2), p. 153. doi: 10.2307/1403610.

Schwarz, N. and Strack, F. (1999) 'Reports of subjective well-being: Judgmental processes and their methodological implications', *Well-being: The foundations of hedonic psychology*, 7, pp. 61–84.

Sen, A. (1987) *On Ethics and Economics*. Oxford.

Sentience Insitute (2018) *Sentience Institute US Factory Farming Estimates - Google Sheets*. Available at: https://docs.google.com/spreadsheets/d/1iUpRFOPmAE5IO4hO4PyS4MP_kHzkuM_-soqAyVNQcJc/edit#gid=0 (Accessed: 18 June 2018).

Singer, P. (1972) 'Famine, Affluence, and Morality', *Philosophy & Public Affairs*. Wiley, 1(3), pp. 229–243.

Singer, P. (1975) *Animal liberation*, *Animal liberation*. London: Jonathan Cape.

Singer, P. (2006) 'Factory Farming: A Moral Issue', *The Minnesota Daily*, 22 March.

Singer, P. (2009) *The Life You Can Save*. Penguin Random House.

Singer, P. (2015) *The most good you can do: How effective altruism is changing ideas about living ethically*. Text Publishing.

Singer, P. (2019) *The Life You Can Save*. 2nd edn. Penguin Random House.

Singer, P., Kissling, F. and Musinguzi, J. (2018) 'Talking about overpopulation is still taboo. That has to change', *The Washington Post*, 18 June.

Steptoe, A., Wardle, J. and Marmot, M. (2005) 'Positive affect and health-related neuroendocrine, cardiovascular, and inflammatory processes', *Proceedings of the National Academy of Sciences*, 102(18), pp. 6508–6512. doi:

10.1073/pnas.0409174102.

Stevenson, B. and Wolfers, J. (2008) *Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox*. Cambridge, MA. doi: 10.3386/w14282.

Stiglitz, K., Sen, A. and Fitoussi, J. (2009) *Report by the commission on the measurement of economic performance and social progress*.

StrongMinds (2015) *Phase Two Impact Evaluation Report*.

StrongMinds (2018) *StrongMinds Q1 Report 2018*.

Sumner, L. W. (1996) *Welfare, happiness, and ethics*. Clarendon Press.

Taurek, J. M. (1977) 'Should the Numbers Count?', *Philosophy & Public Affairs*. Wiley, pp. 293–316. doi: 10.2307/2264945.

The Life You Can Save (2019) *Best Charities to Donate to*. Available at: https://www.thelifeyoucansave.org/best-charities (Accessed: 8 August 2019).

Tiberius, V. (2006) 'Well-Being: Psychological Research for Philosophers', *Philosophy Compass*. John Wiley & Sons, Ltd (10.1111), 1(5), pp. 493–505. doi: 10.1111/j.1747-9991.2006.00038.x.

Tomasik, B. (2015) 'The Importance of Wild-Animal Suffering', *Relations. Beyond Anthropocentrism*, 3(2), pp. 133–152. doi: 10.7358/rela-2015-002-toma.

Tourangeau, R., Rasinski, K. A. and Bradburn, N. (1991) 'Measuring Happiness in Surveys: A Test of the Subtraction Hypothesis', *Public Opinion Quarterly*. Narnia, 55(2), p. 255. doi: 10.1086/269256.

Udayashankar, C., Oudeacoumar, P. and Nath, A. (2012) 'Congenital insensitivity to pain and anhidrosis: A case report from South India', *Indian Journal of Dermatology*.

Wolters Kluwer -- Medknow Publications, 57(6), p. 503. doi: 10.4103/0019-5154.103080.

United Nation Department of Economic and Social Affairs (2017) *World Population Prospects: The 2017 Revision, Key Findings and Advance Tables. ESA/P/WP/248*.

US Poultry (2014) *Poultry Insight: What is an AFO and What is a CAFO? - YouTube*. Available at: https://www.youtube.com/watch?v=aq662XeMe3g (Accessed: 18 June 2018).

Vegetarian Resource Group (2016) *How Many Adults in the U.S. are Vegetarian and Vegan?* Available at: https://www.vrg.org/nutshell/Polls/2016_adults_veg.htm (Accessed: 18 June 2018).

Voorhoeve, A. (2014) 'How Should We Aggregate Competing Claims?', *Ethics*. University of Chicago PressChicago, IL, 125(1), pp. 64–87. doi: 10.1086/677022.

Walen, A. (2016) 'Retributive Justice', *Stanford Encyclopedea of Philosophy*.

Weathers, S. (2016) *The Meat Eater Problem: Developing and EA Response*, *Effective Altruism Forum*. Available at: https://ea.greaterwrong.com/posts/gA57ThbaS3znib242/the-meat-eater-problem-developing-an-ea-response (Accessed: 5 July 2019).

Wenar, L. (2015) *Blood oil*. OUP.

Wexler, A. and Reiner, P. B. (2019) 'Oversight of direct-to-consumer neurotechnologies', *Science*. American Association for the Advancement of Science, 363(6424), pp. 234–235. doi: 10.1126/science.aav0223.

WHO (2017) *About social determinants of health*, *WHO*. World Health Organization. doi: 10.1109/PCT.2007.4538627.

Wiblin, R. (2016) *The Important/Neglected/Tractable framework needs to be applied with care*, *Effective Altruism Forum*.

Wiles, N. J. *et al.* (2016) 'Long-term effectiveness and cost-effectiveness of cognitive behavioural therapy as an adjunct to pharmacotherapy for treatment-resistant depression in primary care: follow-up of the CoBalT randomised controlled trial', *The Lancet Psychiatry*. Elsevier, 3(2), pp. 137–144. doi: 10.1016/S2215-0366(15)00495-2.

Wilson, T. D. and Gilbert, D. T. (2005) 'Affective Forecasting', *Current Directions in Psychological Science*. SAGE PublicationsSage CA: Los Angeles, CA, 14(3), pp. 131–134. doi: 10.1111/j.0963-7214.2005.00355.x.