A philosophical review of Open Philanthropy's Cause Prioritisation Framework

Michael Plant

July 2022

Happier Lives Institute

Contents

Summary

In this post, I undertake a philosophical review of Open Philanthropy's Global Health and Wellbeing Cause Prioritisation Framework, the method they use to compare the value of different outcomes; in practice, the framework focuses on the relative value of just two outcomes, increasing income and adding years of life. This is a 'red-team' exercise, partially inspired by the EA forum contests on blogging and on critiques. My aim is to constructively critique something that is already good with the aim of improving it.

Here's the prospectus. I sketch the various philosophical debates such a framework must take a position on. I note some of the options in each case and how choosing one rather than another may relatively alter the priorities. I attempt to identify the choices that the Open Philanthropy (OP) framework makes and conclude it is substantially unclear which positions they take or which arguments they accept. Specifically, to make the central trade-off between health and income OP implicitly appeals to different, opposing theories of wellbeing. Further, it is not obvious what assumptions the framework makes about population ethics and the badness of death, nor are any reasons given on these topics.

I make three main recommendations:

1. I suggest OP's framework would be greatly improved by clarifying and developing their stance on these philosophical issues.

2. Given OP's commitment to worldview diversification (putting significant resources behind every worldview they find plausible), I argue that OP is pushed to adopt more worldviews within the scope of its Global Health and Wellbeing Framework than the single one it adopts currently.

3. I suggest how OP might, in practice, expand their set of worldviews.

One crucial worldview I argue should be included - but which OP is currently sceptical of - is the use of subjective wellbeing measures (self-reports of happiness and life satisfaction) as a unified measure of value. I have previously advocated for a 'SWB approach' to cause prioritisation (see, for example, here and here). What's novel here is the claim that using a SWB approach would be consistent with, indeed possibly even required by, OP's worldview diversification approach. The simple thought is that it is plausible to think both that (1) happiness and life satisfaction matter morally and (2) people's own ratings of their happiness and satisfaction are a credible way of measuring those feelings.

While this discussion focuses on OP's framework, the issues I discuss here are very general and apply to any method which involves comparing the value of different outcomes. I consider OP's framework as it is a prominent and concrete means of doing this.

1. Background

Open Philanthropy's mission is "to help others as much as we can with the resources available to us". Since 2014, it has directed over $1.5 billion in grants and in 2022-23 it plans to give an additional $1 billion to evidence-based charities working in global health and development.

Open Philanthropy (OP) has done an enormous amount of good. In particular, I want to praise them for the transparent way they communicate their reasoning so that others may learn from, use, and suggest improvements to it. For OP to do this is a public service, one that is almost unique in the wider world of philanthropic actors.

In November 2021, OP published a blog post setting out an updated framework for making trade-offs between different types of interventions in their global health and wellbeing portfolio.[1] In January 2022, Alexander Berger (co-CEO) and Peter Favaloro (Research Fellow) presented the framework at an online seminar hosted by the Centre for Global Development (CGDev). In practice, the framework sets out to establish the relative quantitative value of two key outcomes, improvements to income and health (more details later).

In this post, I provide a philosophical review of the framework and suggest how it could be improved. Doing good well is difficult. I am very grateful for OP in explicitly presenting a framework that is clear enough it can be evaluated and built on. Before we get to the review itself, a few preliminary comments are in order.

Why focus on OP's framework? Besides their own substantial resources and openness to debate, they are proposing this framework as something others could use to achieve maximum impact. OP is not saying "this is how *we* plan to assess impact, but there are no good reasons for others to use it". Hence, it seems worthwhile to assess the approach on its merits.

I recommend reading OP's blog post and/or watching the seminar before reading this, but these are not required to understand what follows.

Readers should be aware that OP takes a worldview diversification approach to its grantmaking. As they put it, this entails "putting significant resources behind *each* worldview that we find highly plausible". By 'worldview' I understand OP to mean 'a set of philosophical assumptions' - I'm not sure how else to conceptualise the term.

I assume that most readers will expect that making different philosophical assumptions leads to different practical implications. In what follows, I show that our philosophical choices within each

[1] The global health and wellbeing approach can be contrasted with the longtermist approach to doing the most good. The distinction is somewhat fuzzy, but the gist is the latter aims to maximise impact over the very long run, i.e. over thousands of years, and the former aims to 'do good now'. See OP's blogpost for discussion.

debate can lead to meaningful practical differences.[2] Curiously, OP appear to be sceptical of this for reasons I do not fully understand; I discuss this issue in a footnote.[3]

OP's prioritisation framework is very similar to that of GiveWell. Hence, what I say here will apply, to some extent, to GiveWell too. However, I focus on OP because they have more recently and explicitly updated their methods.

2. The philosophical review

There are three main philosophical issues that a framework for determining the value of outcomes must, implicitly or explicitly, take a stand on - theories of wellbeing, the badness of death, and population ethics.[4] The following sections introduce each debate in turn.

2.1 Theories of wellbeing

In the blog post and seminar, there are many references to value, wellbeing, welfare, and impact. As far as I can see, these terms are not further defined. How might we understand them?
In philosophy, wellbeing refers to what makes life go well, what is ultimately good for a person (*wellbeing* and *welfare* are synonymous). There are three standard accounts of wellbeing: *hedonism*, where wellbeing consists in *happiness* (a positive balance of pleasant over unpleasant experiences); *desire satisfaction theories*, where wellbeing consists in having your desires met; and *the objective list*, where wellbeing consists in various objective goods such as knowledge, love, and achievement but may also include happiness and/or satisfied desires.

All plausible moral theories accept that wellbeing has intrinsic value i.e. wellbeing is of *ultimate* value, rather than being merely *instrumentally* valuable (valuable because it achieves some further goal). *Welfarism* is the view that wellbeing is the only intrinsic good. *Non-welfarism* is the view that wellbeing is not the only intrinsic good - other things, such as equality or justice, are valuable too. The claim that wellbeing has instrinsic value should not be, but often is, confused for *egoism*,

[2] If more precision is required, I mean that, within each philosophical debate, switching from one view to another will alter the relative value of some pair of choices by a factor of at least two.

[3] Footnote 50 of OP's Technical Update states "Our analysis tends to find that picking the wrong moral weight only means sacrificing 2-5% of the good we could do [...]" and links to a spreadsheet. I confess I cannot establish what the argument is from eyeballing the cells in the spreadsheet. I assume there's something about diminishing marginal returns for very large donors, but I don't know what else is going on. That said, if OP thinks that taking different views on ethics genuinely make almost no practical difference, that would be an astonishing result, one I would encourage them to explain and defend at greater length. I assume the thought is more like "given some restrictions on what our moral views can be, we find it makes little difference which ones we pick". Then, of course, the concern is that the choice of moral views has been unduly restricted.

[4] I assume here we are only interested in the value of outcomes, that is, axiological issues. I am not interested in normative issues, that is, those related to what we ought to do.

the moral view that individuals ought only care about their own wellbeing. To illustrate by contrast, *utilitarianism* is a moral theory that holds wellbeing is the only intrinsic good, but individuals ought to promote the welfare of everyone (not just themselves). OP seem exclusively interested in promoted wellbeing, so that's all I focus on here.

Which account of wellbeing is endorsed by the framework? The framework takes changes to two goods, income and health, and then uses various different methods to determine their value compared to each other, without stating what ultimately matters (we will come to these methods shortly). More specifically, OP says:

> *In short, we use a logarithmic model of the utility of income, so a 1% change in income is worth the same to everyone, and a dollar of income is worth 100x more to someone who has 100x less. We measure philanthropic impact in units of the welfare gained by giving a dollar to someone with an annual income of $50,000, which was roughly US GDP per capita when we adopted this framework. Under the logarithmic model, this means we value increasing 100 people's income by 1% (i.e. a total of 1 natural log unit increase in income) at $50,000. We have previously also valued averting a disability-adjusted life year (DALY; roughly, a year of healthy life lost) at $50,000, so we valued increasing income by one natural-log unit as equal to averting 1 DALY.*

One key change in the update being that:

> *We're doubling the weight on health relative to income in low-income settings, so we will now value a DALY at 2 natural log units of income or $100,000.*

One issue with this approach is that income and health are not *intrinsically* valuable for us. I think most people would agree that they are *instrumentally* valuable, they are good for us only because they increase our wellbeing. Given this, the only theoretically coherent way to justify their relative value is to state what you think ultimately matters e.g. happiness, and then explain that income and health are valuable *because and to the extent that* they increase whatever ultimately matters. Indeed, OP seems to recognise this, because they go on to offer further explanation for how they reached these relative values (I come to these explanations shortly). If you argue X is good only because it results in Y, then you cannot believe X is what ultimately matters.

The further issue with just focusing on these two outcomes, health and income, is that there are other things that affect our wellbeing, such as war, loneliness, unemployment, crime, living in a democracy, etc. whose value is not entirely reducible to effects on health or income. OP is aware of this issue and concerned about it (see Alexander Berger's answer around 1:14:00 in the seminar).[5]

[5] Or, at least, this is an issue if you conceive of health as being the absence of infirmity or disease. The WHO, curiously, states that "Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity". In which case, everything will affect health because health = wellbeing but then, of course, this raises the question, "what is wellbeing?"

Therefore, to do cause prioritisation well, we need a principled overarching concept - and measure - of value that provides a method for comparing across the full range of outcomes.

Let's now take a closer look at how OP justifies their relative weights and how rethinking this can lead to different priorities and help us to do good better.

2.1.1 Income and quality of life

In the seminar (08:44), Peter Favaloro explains OP's choice of a logarithmic relationship between income and utility by pointing out that the subjective wellbeing (SWB) evidence supports this conclusion. That is, when you survey people, a 10% increase in income has the same effect on SWB at any level of income. Hence, OP are making an implicit claim here that SWB matters.

This raises two questions. First, how does SWB fit with the three theories of wellbeing I mentioned earlier? Second, if SWB does capture something of deep moral importance, why not measure the effects of everything in terms of subjective wellbeing? If we can measure wellbeing directly, assessing value in terms of health and income would be regressive, not least when those aren't the only two things we're interested in.

Regarding the first question, there are two main types of subjective wellbeing measures (OECD 2013). There are *hedonic* measures of experience and emotion in the moment and *evaluative* measures such as overall life satisfaction. If you're a hedonist, that is, you think wellbeing consists in happiness, you'll care most about hedonic measures, as the name suggests. However, evaluative measures presumably have some value on hedonism - how happy we feel *about* our lives will be related to how happy we are *during* our lives.

Evaluative measures are more closely linked to the desire satisfaction account of wellbeing because they capture how you think your life is going compared to how you'd like it to be going (or, at least, so I argue here). However, desire satisfaction theorists are also likely to care about happiness - it is one of the things we tend to want - and so would take *some* interest in the hedonic measures too.

If you prefer the objective list theory of wellbeing, how much you care about the hedonic and evaluative components of SWB depends on your exact theory. However, it would be surprising if people's happiness and/or life satisfaction played no part, either intrinsically or instrumentally, on any objective list theory of wellbeing.

To sum that all up then, depending on the particular SWB measure and your preferred theory of wellbeing, SWB data is either going to capture *some or all* of wellbeing.[6] Happiness and life satisfaction are relevant to *any* plausible theory of wellbeing. No matter how much we want to

[6] This assumes that SWB surveys are 'valid', that is, they capture what they intend to capture. For this, see OECD (2013) and references therein.

incorporate them in our cause prioritisation, we will still want to use the most accurate, practical measures of them.

And what about the second question, that of using SWB as a unifying metric? When asked about this in the seminar, Alexander Berger provided a detailed response (1:15:21) which I paraphrase here for length and clarity (you can read the full quote in the footnotes).[7] Berger comments that an attractive feature of SWB is that it *could* be a unified outcome variable. However, there are different SWB questions. Although life satisfaction seems to correlate with what we assume matters, it is "sort of a random subjective question" and lacks "a deeply privileged normative status". He goes on to say, "if you're a hedonic utilitarian - which mostly we [OP] are not - you'd be more interested in the other kind of happiness."

This response raises several questions that require further clarification:

1. Given that hedonism is a worldview that many take seriously, doesn't it follow that OP, by its self-stated approach to worldview diversification, should commit to putting some non-trivial resources towards it?

2. If life satisfaction scores lack a "privileged normative status", why appeal to them to justify the choice of a logarithmic model of income?

3. Considering that life satisfaction looks like a credible measure of wellbeing on the desire satisfaction theory of wellbeing (see my discussion here), shouldn't we say it does have a "privileged normative status" on that account?

4. If one accepts that life satisfaction is a good measure on desire theories, but it nevertheless lacks a "privileged normative status", this suggests OP *also* reject desire satisfaction theories. If they reject both hedonism *and* desire theories, that leaves only the objective list. Do they reject that too? I hope not, because, if they did that would mean rejecting *all* of the available options!

[7] Transcribed from talk (lightly edited for clarity): "One natural virtue of subjective wellbeing (SWB) as a unified outcome is that it could be a space on to which you could project all of these different things. And that sort of makes it attractive. As Charles Kenny pointed out in the chat, a lot of things don't impact SWB quite as much as you might think and different SWB measures get you totally different answers for what is valuable. And so, in Peter's chart, that was a life satisfaction (LS) question as opposed to a hedonic, 'how happy do you feel in this moment' type question. The LS questions tend to be more correlated with the other kinds of things we think of as good things in life that people would naturally want to maximize and make go up and to the right and so I think that makes you think that maybe LS is a good kind of outcome but fundamentally LS is sort of a random subjective question to ask somebody about their life. I don't see why it has a deeply privileged normative status. If you're a hedonic utilitarian - which mostly we're not - you'd be more interested in the other kind of happiness -- and that pushes you in unusual ways that may make you care way more about animals. And so I think this becomes philosophically extremely complicated territory very quickly. SWB has some nice virtues as a statistical construct but...as a starting point, i don't see it as a super compelling normative unit."

I assume that OP, like many, is not strongly convinced by any one of the three theories of wellbeing but has some non-trivial credence in each. If this is the case, the thing for OP to do, given their commitment to worldview diversification, is to take each theory seriously. What would that involve? Well, you select a suitable measure for each theory of wellbeing and identify what the priorities would be in each case. Presumably, the way worldview diversification works is that you divide your resources in proportion to your credence for each worldview. So, if you gave a 20% chance to hedonism being true, you'd put 20% of the resources in the hedonism 'bucket' and so on.

I've already noted above that hedonic and evaluative subjective wellbeing questions seem to map reasonably well onto hedonism and desire theories of wellbeing, respectively. It's less obvious how to measure the objective list as there are different versions.

It seems that OP do not find measures of SWB entirely implausible - why else appeal to life satisfaction to justify the logarithmic relationship to income? In this case, the obvious thought is to use those measures directly, rather than rely on income and health metrics instead.

What difference might it make to take a subjective wellbeing aproach to cause prioritisation, rather than OP's current approach of (log) income and health metrics? Will we find new priorities? I think that we will - indeed, perhaps we already have.

In the first research of its kind, my colleagues and I at the Happier Lives Institute conducted two meta-analyses to compare the cost-effectiveness, in low-income countries, of providing psychotherapy to those diagnosed with depression compared to giving cash transfers to very poor families. We did this in terms of subjective measures of wellbeing and found that therapy is 9x more cost-effective. You can read an overview of our research here. To the best of our knowledge, these are the first two meta-analyses assessing the cost-effectiveness of anything in terms of its impact on subjective wellbeing. Meta-analyses of the effect of interventions are common, but not of their total effects (i.e. effects across time) or their cost-effectiveness.

This result is striking and surprising to many. OP and the wider effective altruism community considers cash transfers a 'good buy' and use them to 'set the bar' for judging what to fund. GiveWell, a sister organisation to OP, aim to fund interventions that are 7-8x more cost-effective than cash transfers (see OP blogpost). Up to now, mental health interventions have not been considered cost-effective enough to clear this bar - but no one had attempted a comparison in terms of SWB. Our analysis, using this different (and, we claim, preferable) methodology, estimates that psychotherapy would clear this bar.

This raises the question of why using the SWB approach might reveal different priorities. The simple answer here is that if we aim to increase happiness and/or life satisfaction, but we *don't* use the observed evidence from people's self-reports, then we have to rely on educated guesses of how much we *think* different things will affect SWB. Unfortunately, research indicates our predictions

are systematically flawed; what we think will make us happy is not what actually does. To quote Gilbert and Wilson (2007):

> *Mental simulations are deficient because they are based on a small number of memories, they omit large numbers of features, they do not sustain themselves over time, and they lack context. Compared to sensory perceptions, mental simulations are mere cardboard cut-outs of reality.*

For instance, our affective forecasts (predictions of how others, or our later selves, feel) are subject to *focusing illusions*, where we overweight the importance of easy-to-visualise details, and *immune neglect*, where we forget that we will adapt to some things and not others, amongst other biases (Gilbert and Wilson 2007). This suggests our intuitive cause prioritisation attempts will overweight visible suffering (such as poverty) and underweight invisible, ongoing misery (such as mental illness or chronic pain). In broad terms, I take the implication of this affective forecasting research to be that we shouldn't trust our intuitions and we really need to rely on the data.

It's worth pointing out that QALYs and DALYs, the standard health metrics that OP, GiveWell, and others have relied on in their cause prioritsation framework, are likely to be misleading because they rely on individuals' assessments of how bad they *expect* various health conditions would be, not on observations of how much those health conditions alter the subjective wellbeing of those who have them (Dolan and Metcalfe, 2012).

Based on all this, taking the SWB approach seriously as a worldview might lead OP (and others) to quite different conclusions about what the priorities are.

I should note the further possibility that using different measures of subjective wellbeing *could* push us in different directions; the best thing for improving happiness may be different from the best thing for increasing life satisfaction. Investigating this requires extra work. However, this is the sort of thing we would like to find out about if we take a worldview diversification approach.

2.1.2 Health and quantity of life

In their updated framework, OP states:

> *We're doubling the weight on health relative to income in low-income settings, so we will now value a DALY at 2 natural log units of income or $100,000.*

How is this relative weight reached? OP appeals to two approaches. First, the value of a statistical life (VSL) approach - which is mainstream in economics - puts a price tag on a year of life. This can be done using *stated preferences*, e.g. "would you rather reduce your risk of dying this year by 1%, or increase your income this year by $500?" or by *revealed preferences*, e.g. observing how much more people get paid to work in more dangerous jobs and calculating the value of a life from that. Second, using a SWB approach to compare improving to extending lives (I expand on this shortly).

Looking at these two approaches, OP conclude:

> *[t]here is huge uncertainty/disagreement across and between lines of evidence [...] and any given choice of ultimate valuations seems fairly arbitrary, so we also prefer a visibly rough/round number that reflects the arbitrariness/uncertainty.*

I would be sympathetic to OP's conclusion *if* the two methods were equally informative and gave roughly the same answers. However, when we dig a bit deeper into the philosophy, neither seems to be the case.

The issue with the VSL approach is that it seems to require taking a particular worldview very seriously, namely, that we should give (non-trivial) weight to people's ill-informed desires. I suspect that, on reflection, few people would give much credence to this view.

Regarding the SWB approach, I disagree with the way that OP has interpreted the meaning of SWB scales. It's not at all clear that we should peg two extra years of life as equivalent in value to $100,000.

As a result, I suggest that OP should return to the philosophical drawing board and get clearer on what they think really matters.

The VSL approach

The value of a statistical life (VSL) approach suffers from serious and long-acknowledged problems (Bronsteen, Buccafusco, and Masur, 2014). People are not very good at thinking about probabilities. We are also not very good at predicting how something will make us feel in the future (as noted already).[8]

Where does the VSL approach fit into the three theories of wellbeing? It seems it only makes sense to take this approach seriously if one endorses the desire satisfaction theory of wellbeing. To see this, note that VSL is based on people's observed or hypothetical choices, not on how happy they are. If one wanted to know how they actually felt, one would use SWB data instead.

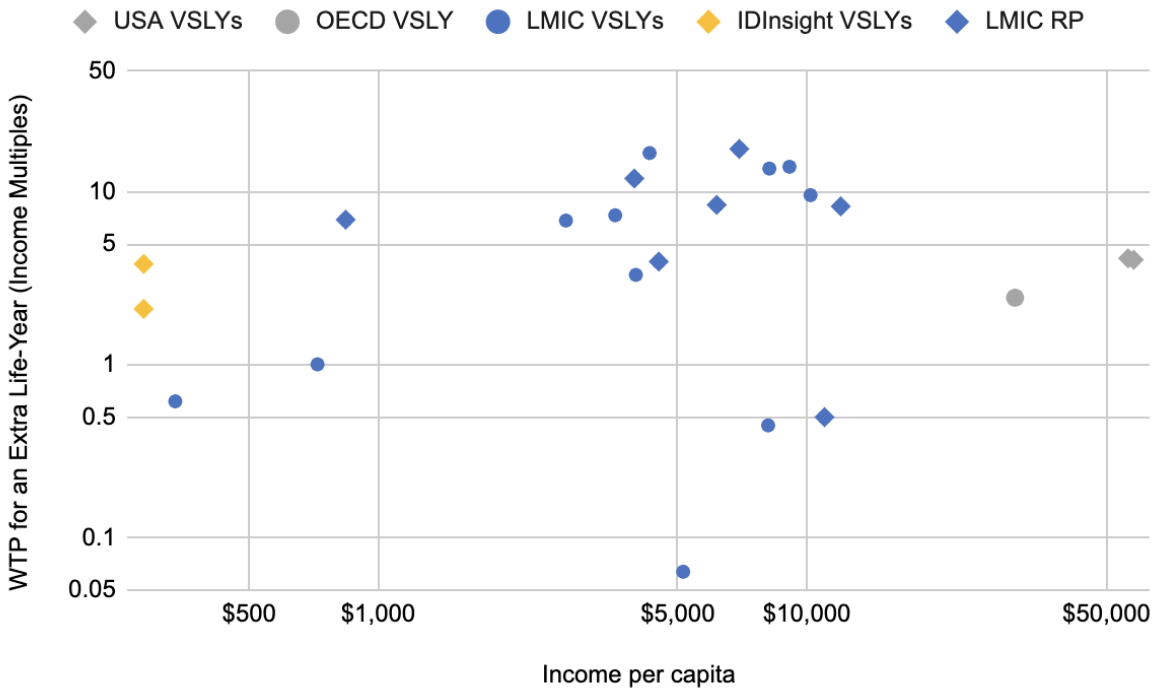
A question all desire satisfaction theories must answer is, which desires count? Amongst those who advocate for desire satisfaction theories, it is commonly (although not universally) argued that we should ignore people's 'defective desires', such as those that are ill-informed or irrational - see Heathwood, (2005, p489) for a list of several philosophers who take defective desires to be a decisive objection to 'actualist' desire theories. Yet, worryingly, it seems that the desires elicited by VSL are of this type.

[8] For a fuller analysis of how global priorities research should account for affective forecasting, see HLI's 2022 report.

Consider, as one example, the beneficiary preferences study conducted by IDinsight that OP appeals to in order to support their own choice of moral weights (see GiveWell 2019 for a write-up). The survey asked potential beneficiaries of GiveWell top charities how they would trade-off changes to income against lives saved. OP notes that "these respondents would value an extra year of life expectancy as much as 2.1-3.8 years of income which we [OP] can interpret as 2.1-3.8 units of log-income." Yet, as GiveWell notes, "About 34% of respondents were able to answer the following question correctly: Which risk of death is larger: 1 in 100 or 2 in 1,000?" (GiveWell, 2019). Equally, "In the choice experiment framing, 38% of respondents did not believe that the equivalent of $10 million in cash transfers would be more beneficial than saving a life" (GiveWell, 2019). If we think it is people's *informed preferences* that matter, we shouldn't ignore these methodological worries and take the answers at face value. Instead, we should place little, if any, weight, on such questions (or, at least, answers from individuals who fail relevant tests of comprehension).

Alternatively, we might decide that we do want to take people's preferences seriously, whether they are well- or ill-informed. To do this, one must not only accept (implicitly) a particular view of wellbeing - the desire satisfaction theory - but a particular *version of that view* that is out of favour even amongst desire theorists. We should know which bullets we are biting.

Suppose, nevertheless, that we did want to take the VSL literature at face value. What would we conclude? Unfortunately, the range of estimates is very wide - an issue which OP readily acknowledges and demonstrates in this graph, which I reproduce from their blog post. The y-axis represents 'willingness to pay' (WTP) for an extra year of life in terms of multiples of annual income. A WTP of 2 would mean you'd give up two years of your salary to live one year longer.



As you can see, the WTP numbers range from less than 0.1 to more than 25, and they are scattered all over the place. OP concludes that an extra year of life is worth two years of income but this choice seems substantially arbitrary. Hence, the use of the VSL approach in general, and the 2x choice in particular, merits further defence and explanation.[9]

The subjective wellbeing approach

OP also appeals to the SWB literature for their health-to-income multiple. They conclude, drawing on a graph from Stevenson and Wolfers (2013), that:

> *The chart suggests that, for someone near the global poverty line, you'd have to increase income by roughly 64x in order to get twice as much life satisfaction at any moment. You could conclude that a 64x increase in income (for one year) is worth as much life satisfaction as an extra year of life. In that case, a logarithmic utility function suggests valuing a DALY at roughly 4 units of log-income.*

This doubling is supposed to occur as the result of someone going from 4/10 on the scale to 8/10. This relies on the assumption that 0/10 is the 'neutral point' on the scale, where the person has zero wellbeing - there life is overall neither good nor bad for them.[10] However, determining the location of the neutral point is an open question that I've written about before. This is an

[9] I recognise that OP appeal to various other pieces of evidence, saying that "the best US VSL studies we know of" value a year of life expectancy 2.5 to 4 times more than annual income. However, I would like to know what it is that makes these the "best" studies, such that we should still take them seriously in light of the worries raised above.

[10] More technically, we can say that this assumes we have a ratio scale where 0/10 is the non-arbitrary zero point.

important methodological issue on which, as far as I am aware, there is no published work. I am working on a paper on this with some colleagues and hope to set out some more detailed thoughts soon. For now, I hope it will suffice to indicate that this is an open question.

A neutral point of 0/10 on a life satisfaction scale implies that life can never be bad - it can only be neutral or good - as you cannot give an answer lower than 0 on the scale.

An alternative position is to take 5/10, the middle of the scale, as the neutral point on the grounds it indicates the person is overall neither satisfied nor dissatisfied with life. In this case, going from a 4/10 to an 8/10 does not represent a doubling - a move from being somewhat satisfied to twice as satisfied. It is a far more dramatic improvement: the person has gone from a life at -1 units of wellbeing, i.e. a life with negative wellbeing, to +3 units, a life with substantially positive wellbeing.

However, a neutral point of 5/10 also leads to the conclusion that anyone with a LS score less than 5 is living a life of negative wellbeing, which suggests that the framework should count it as bad, rather than good, to extend their life.

A third option is to somehow assign a neutral point between 0/10 and 5/10 but this seems dangerously *ad hoc*. You could also assign the neutral point at 10/10, but this would imply that extending lives is always bad (for the individual).

It's worth pointing out that many of those whose lives are saved by the interventions that OP funds, such as anti-malaria bednets, will have a life satisfaction score below the neutral point, unless we set it at, or near to, 0/10. IDinsight's aforementioned beneficiary preferences survey has a SWB question and found those surveyed in Kenya had an average life satisfaction score of 2.3/10.

That saving such lives is, in fact, bad (for the individual) is initially highly counterintuitive, but is less so on further reflection. Here are some options to consider:

1. Perhaps such lives have negative wellbeing now, but will improve in future, so saving them is overall good. In which case we need to predict how their lives will improve and how far above the neutral point the lives will be.

2. We might conclude we are using the wrong measure of wellbeing - such people may be dissatisfied, but happiness is what matters, and we think they are happy and so saving their lives is good. In which case we need to measure their happiness (and predict their future happiness).

3. If we think those living in extreme poverty really suffer - which, after all, is often the motivation for providing aid - wouldn't it, in fact, be odd to simultaneously conclude that few, if any, have lives of negative wellbeing.

4. Perhaps the issue is that we have the wrong intervention in mind. We generally accept that we shouldn't try to prolong some lives, such as those in agony with cancer at the end of life. What we should do for those people is try to improve the quality, not quantity, of their lives. Hence, if we think people whose lives we might try to save are truly miserable, then we should consider how best to also improve the quality of their lives. We can then consider how cost-effective doing that is compared to other life-improving interventions to see what does the most good.

If saving lives is bad, then of course OP's claim that an extra year of life is good - and, more specifically, of the same value of 2 log-units of income - would be mistaken.

At this point, I should address the fact that OP seem to think that we should *always* assign a positive value to saving a year of life, and this should be the same value regardless of the level of wellbeing the saved person has. OP states that, unless they make this assumption, it has the intuitively unacceptable implication that saving lives in richer countries would, other things being equal, be more valuable on the grounds that such people are richer and so better off (see section 5.2 in the blog post). There is potentially a lot to say here, but I will be brief.

First, the objection probably relates to a sense of fairness. It's unfair to benefit someone simply because they are lucky enough to be better off. But OP (and others) tend to ignore fairness; the aim is just to do the most good.

Second, this indicates a tension in OP's thinking. OP holds it is better, all things equal, to save someone in their 20s than in their 70s (I discuss this further in Section 2.2). Presumably, this is because the benefit to the younger person is greater. But, don't happier people gain a greater benefit from an extra year of life than less happy people? If so, how can it be consistent to conclude we should account for *quantity* when assessing the value of saving lives, but not *quality*? I would welcome OP to expand on these concerns.

My aim is not to settle these issues here. As noted, I plan to return to (some of) them in further work. My aim is simply to highlight the issues in order to show that OP's conclusion that the life satisfaction evidence "suggests valuing a DALY at roughly four units of log-income" is not obviously supported. I am showing more philosophy is needed, not claiming I can provide here.

2.2 Theories of the badness of death

An additional philosophical choice for cause prioritisation frameworks is how to compare the badness of death for different people. OP comments:

> *We're also switching our approach to be more consistent with GiveWell's framework in how we translate deaths into DALYs. GiveWell assigns moral weights to deaths at various ages, rather than to DALYs [...] The resulting model cares about child mortality more than adult mortality, but not by as much as remaining-population-life-expectancy would suggest.*

To explain what's going on here, it may help to know this is making implicit claims about what, in philosophy, are known as *theories of the badness of death*. I've commented on this previously here. For a good edited compilation on this topic, see Gamlund and Solberg (2019).
Perhaps the standard view of the badness of death is *depriviationism*, which states that the badness of death consists in the wellbeing the person would have had, had they lived. On this view, it's more important to save children than adults, all else equal, because children have more wellbeing to lose.

Some people have an alternative view that saving adults is more valuable than saving children. Children are not fully developed, they do not have a strong psychological connection to their future selves, nor do they have as many interests that will be frustrated if they die. The view in philosophical literature that captures this intuition is called the *time-relative interest account* (TRIA).

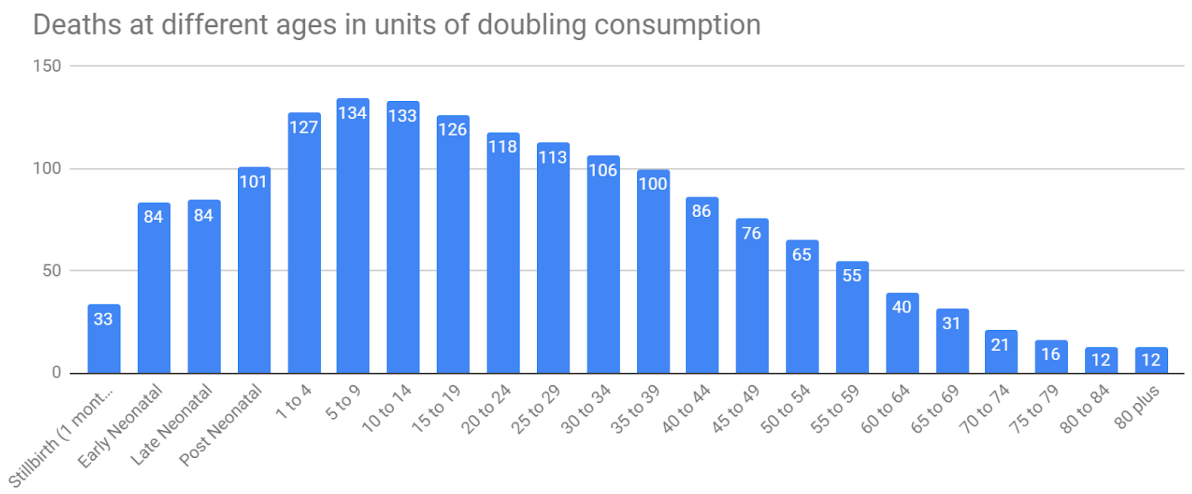
A third view is *Epicureanism*, named after the ancient Greek philosopher Epicurus, on which death is not bad for us and so there is no value in living longer rather than shorter.[11]

Clearly, these three views have different implications. Depriviationism holds that saving children is better than saving adults, TRIA holds the reverse to be correct, and Epicureanism implies we should focus on improving lives rather than saving lives. It's an open question exactly how TRIA should be operationalised, so it's unclear exactly how the numbers might work out. A first, crude attempt by myself and colleagues, using SWB data, found the value ratio of averting the death of an under-5 to doubling consumption of one person for one year is 154:1 on deprivationism and 33:1 on TRIA. Hence, the choice of view is potentially quite substantial, if switching from one view to another makes saving lives six time more (or less) valuable.

What is OP's view on the badness of death? They say "the resulting model cares about child mortality more than adult mortality, but not by as much as remaining-population-life-expectancy would suggest" which indicates they are rejecting depriviationism in favour of TRIA.

What is their reasoning? OP explains they are "going to defer to GiveWell and start to use the number of DALYs that would be implied by extrapolating their moral weights". The figure below shows how GiveWell weight the badness of deaths at different ages.

[11] Perhaps a better way to characterise the view is that death and life are not comparable in value for someone - death is neither better for, worse for, or equally good as, existence.



Where do GiveWell's numbers come from? GiveWell has written about it here. In short, they use three sets of information that assign different weights to each: donor preferences for how to do the trade-off, the intuitions of GiveWell staff, and a standard 'years of life lost' approach - in other words, a deprivationist approach. They state the relative weights are assigned as follows:

- 60% weight on donor responses
- 10% weight on James Snowden's 2018 weights (as a proxy for 2018 GiveWell staff)
- 30% weight on YLLs (both as a commonly-used metric itself and as a proxy for the IDinsight survey)

GiveWell state, "donor responses received the majority of the weight primarily because in our view, that curve represents the most plausible set of weights". But they do not explain *why* these are the most plausible set of weights. On what grounds are the donor preferences the most plausible weights? Presently, OP defer to GiveWell, and GiveWell defer to (or agree with) their donors but without explaining why.

The philosophical literature is rich with arguments for and against each of the views on the badness of death (again, Gamlund and Solberg, 2019 is a good overview). We should engage with those arguments, rather than simply polling people. At least, we should bring out the arguments first, ensure people have engaged with and understood them, then poll them.[12]

As before, if OP is taking a worldview diversification approach, they do not need to go 'all-in' on a single philosophical view. Instead, they could divide up their resources across deprivationism, TRIA, and Epicureanism in accordance with their credence in each view. The current approach

[12] Because academic philosophers observe and experience that seriously engaging in philosophical discussion often leads people to substantially revise their intuitions, they tend to put little weight on people's "untutored intuitions" and far more on those that are tutored.

taken by OP (and GiveWell) is to congeal these different assumptions into a single, combined assumption. Doing this seems inconsistent with the worldview diversification approach.[13]

2.3 Population ethics

The third philosophilcal issue that OP's framework must take a position on is population ethics, which is, roughly, how to think about the value of future people. OP says:

> *We also haven't reached any settled thoughts on the impact of population ethics or the second-order consequences of saving a life (e.g., on economic or population growth) on how to translate between deaths and DALYs.*

Population ethics is a notoriously tricky topic of philosophy, with discussants widely accepting that all of the options have counterintuitive implications. See Greaves, 2017 for clear, brief overview of the field. Hence, I have real sympathy with OP wanting to be non-committal on the issue. The problem is that trying not to take a view on the subject still involves, implicitly, taking a view. And it might matter substantially which view we take.[14]

To illustrate, the (mathematically) simplest view in population ethics is *totalism* (or *the total view*) where the value of an outcome is the sum of wellbeing of everyone in it - past, present, and future. So, totalists just want the sum of wellbeing to be as large as possible and believe that adding more happy lives to the world is good.

An alternative to *totalism* is the family of *person-affecting views* that hold, in slogan form, "morality is about making people happy, not making happy people" (a paraphrase of Narveson, 1973). One common line of thinking that leads to this conclusion is that there is no value in adding happy lives because no one is made better off by being born, and an outcome can't be better if it's not better for anyone. So our moral concerns are restricted to whichever people will exist whatever we do.

I am moving very briskly across this philosophical terrain here - there are many other views and lurking issues, but here is not the place to raise or discuss them. Again, Greaves (2017) is a good place to start.

You might think you can avoid this choice if you have, for whatever reason, decided that you are only concerned with improving and saving lives. However, the number of children that women

[13] The standard approach to moral uncertainty in academic philosophy is to 'maximise expected choiceworthiness', where we make intertheoretic comparisons between different views to decide what we have most reason to do (see Bykvist 2017 for an overview). But worldview diversification appears to be an alternative approach - if one liked maximise expected choiceworthiness, it's unclear why you would practice worldview diversification.

[14] The 2017 GiveWell blog post that OP hyperlinks to is responding to a 2016 blog post where I first raised these issues. If OP have not formed a strong view on population ethics in the last five years, that seems all the more reason to engage in worldview diversification.

have is affected by child mortality rates. Work by David Roodman for GiveWell indicates that for every two people that GiveWell's life-saving interventions save, approximantely one fewer person is born (see GiveWell 2017 blog). So, if two lives are saved, the result is one extra lifetime of existence is lived, not two.

Here is where the population ethical questions bite. If one takes a totalist view, this reduction in fertility is, on its face, regrettable. You wanted the universe to have as much wellbeing as possible, and so, learning about the fertility-mortality interaction, you should now conclude your life-saving intervention was about half as cost-effective as you originally thought. By contrast, if one takes a person-affecting view, then one ignores these fertility effects on the grounds they are irrelevant, so life-saving interventions looks relatively better (roughly as good as it would on the totalist view if there were no replacement).[15]

Hence, there is no neutral choice here. If one ignores the fertility effects, then the framework implicitly adopts the person-affecting view. It is not clear which position OP takes on this.

3. Recommendations

OP's cause prioritisation framework aims to set out how to make difficult choices between hard-to-compare things in order to do the most good. It serves as a model for OP's decision-making as well as others who might want to adopt it. In light of what I've said, how could this methodology be adjusted? I make three recommendations.

3.1 Be explicit about the philosophical assumptions

First, I encourage OP to be more explicit about the philosophical underpinning of their approach. This will help to reveal if their practice is consistent with their underlying beliefs, add clarity to what those beliefs are, and reveal whether further clarity may be needed. It will also help others who might use the model to determine where and why they might disagree with OP's approach.

3.2 Adopt additional worldviews

Second, assuming OP intends to stick with its worldview diversification appproach, it seems the natural step is to adopt further worldviews within the global health and wellbeing 'bucket' of resources. OP already hold they should split their resources to reflect different beneficiary groups, roughly, (1) people in the near-term, (2) animals in the near-term, (3) lives in the long-term. Of course, one might object, "But when would we stop? There could be any number of worldviews". This, in some sense, is an easy question to answer. If OP intends to "[put] significant resources

[15] I'm assuming, for simplicity, that individuals' lives have a neutral overall effect on everyone else. This is a controversial assumption.

behind each worldview that we find highly plausible", then they should have one bucket for each worldview they find plausible and no more.

As theory of value is built out of combining answers to several questions, you could have as many worldviews as there are different plausible combinations. Here, we've discussed three candidate answers to theories of wellbeing, three to the badness of death, and two for theories of population ethics. Naively then, that is 3 x 3 x 2 = 18 different worldviews. What complicates this is that there are many more possibilities. I didn't discuss all the possible candidates in each case - there are many possible objective lists, as well as theories in population ethics. What's more, many ethical views admit of degree. On the other hand, matters may be simplified because there will be many assumptions, and combinations of assumptions, that one simply does not find plausible. There is also the practical realisation that many worldviews would reach the same conclusion about what the top priorities are.

3.3 Adopt the following worldviews

My third recommendation concerns the issue of what a worldview-diversifying agent, such as OP, might do in light of all this. As a practical matter, I suggest a middle option of slightly expanding the worldviews with the global health and wellbeing 'bucket' - rather than having a single worldview or trying to accommodate every worldview.

What might that look like, concretely? When it comes to measuring changes to wellbeing, I've already argued that a subjective wellbeing approach should feature somewhere. Hedonism and desire theories are two credible accounts of wellbeing and seem sensibly measured by happiness and life satisfaction surveys, respectively.

Accounting for these somehow would presumably be a complement to, rather than replacement of, OP's current worldview of measuring wellbeing - what we might call their "(log)income and health" approach. That said, I'm not sure exactly how OP's "(log)income and health" worldview differs from a SWB-based one and I would welcome OP to elaborate on their perspective here. I suspect the theoretically cleaner option would be to replace the "(log)income and health" approach altogether, but we shouldn't let the perfect be the enemy of the good.

When it comes to comparing saving lives to improving lives, OP could consider the badness of death from the deprivationist and TRIA perspectives and well as, if they have non-trivial credence in it, Epicureanism.

I would welcome OP clarifying their stance on population ethics and splitting that into (at least) two worldviews, but I'd understand if they wanted to leave that for a further date.

That results in four worldviews (two ways of measuring wellbeing, two views of the badness of death) and resources could be split between them in proportion to the credence that OP has in

each. An extra split, or series of splits, could also be done to account for different views of where the neutral point is. Then money in each 'pot' goes towards whatever seems best by the lights of that pot. Of course, it might be the case that different worldviews nevertheless have the same priorities.

I recognise this is more complicated, but morality is complicated! This complexity, and our uncertainty, already exists and the choice we have is whether to ignore it or account for it. OP's framework is already mathematically sophisticated so taking different worldviews will often just mean re-doing the cost-effectiveness numbers in a spreadsheet. I don't think doing this is enormously complicated. Two colleagues and I identified in this post how cost-effectiveness for life-saving vs life-improving interventions would change under two different accounts of the badness of death.

4. A final request

More broadly, I encourage other actors trying to do the most good to consider making plain to others how different moral assumptions would change their conclusions (of course, many do this already). This seems like a cooperative thing to do in a community where people have different moral views. It has a 'moral trade' feel to it: "I'm going to try to work out what would do the most good by my lights and, seeing as I'm doing this research, I'm going to tell you what I think would be best on yours too".