On the desire to make a difference

Hilary Greaves, William MacAskill, Andreas Mogensen and Teruji Thomas (Global Priorities Institute, University of Oxford)

Global Priorities Institute | September 2022

GPI Working Paper No . 16-2022

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

On the desire to make a difference

Hilary Greaves, William MacAskill, Andreas Mogensen and Teruji Thomas
September 2022

Abstract

True benevolence is, most fundamentally, a desire that the world be better. It is natural and common, however, to frame thinking about benevolence indirectly, in terms of a desire to make a difference to how good the world is. This would be an innocuous shift if desires to make a difference were extensionally equivalent to desires that the world be better. This paper shows that at least on some common ways of making a "desire to make a difference" precise, this extensional equivalence fails. Where it fails, "difference-making preferences" run counter to the ideals of benevolence. In particular, in the context of decision making under uncertainty, coupling a "difference-making" framing in a natural way with risk aversion leads to preferences that violate stochastic dominance, and that lead to a strong form of collective defeat, from the point of view of betterness. Difference-making framings and true benevolence are not strictly mutually inconsistent, but agents seeking to implement true benevolence must take care to avoid the various pitfalls that we outline.

1. Introduction and motivations

Let *benevolence* be a concern to promote the good. It is widely agreed that benevolence is morally laudable: it is a virtue (Hurka 2001). And consequentialists and most deontologists alike agree that there is at least pro tanto reason to promote the good, so that possessing a suitable degree of benevolence disposes one towards right action (Ross 1930:58, Kagan 1991:47).

More precisely, we take it, the benevolent agent tends to *prefer* to *better* promote the good. In its purest form, we will assume, benevolence is a matter of having preferences that track betterness: the purely benevolent agent prefers A to B iff A is better than B.

It is very common, however, to think of one's would-be benevolent activity not directly in terms of the betterness of outcomes, but more directly in terms of *making a difference*. An agent with difference-making preferences, we stipulate, prefers A to B just if, and then because, she makes a bigger (better) *difference* to goodness if A than if B. For example, Singer (2015) writes that at the time of his seminal work on global poverty and animal liberation, his motivation was "thinking about where I could make the *most difference*" (p.118; emphasis in original). Similarly, MacAskill (2015) writes that "effective altruism… is about trying to make the *most difference* you can" (p.15; emphasis in original).

We submit, though, that the ideology of difference-making seems attractive primarily because it is natural to assume that caring about making a difference is extensionally equivalent to caring about the world being better.[1] Insofar as that extensional equivalence holds, asking oneself "how can I make the most difference?" rather than "how can I make the world as good as possible?" is simply an innocuous mechanism for focussing one's thoughts on the factors that are under one's own control. But, as we will show, this extensional equivalence can importantly fail.

Consider, for example, the following decision situations. Both raise pressing real-world practical questions for a would-be philanthropist concerned with promoting the good:

> **Diversification:** Should one donate the whole of one's philanthropic pot to a single charity, or instead split it between multiple charities?
>
> **Cause prioritisation**: Should one fund global poverty alleviation, or mitigation of existential risks?

Anecdotally, many people who face these decision situations and who take themselves to be aiming at "doing the most good" prefer to diversify, and prefer to fund poverty alleviation rather than existential risk mitigation, even in cases in which the opposite choice leads to greater expected value. The rationalisation offered is often in terms of risk aversion. But this is a rationalisation that only makes sense if the risk aversion in question is specifically aversion to risk *in the difference one makes oneself*; the rationalisation does not go through if one is simply risk averse with respect to the total amount of value in the world. And such "difference-making risk aversion", we will argue later in the paper, is in deep tension with the ideals of benevolence. Similar issues apply to the natural combination of ambiguity aversion with a "difference-making" mindset.[2]

One way to put the point is in terms of the distinction between agent-neutral and agent-relative valuations. We normally think that matters of promoting the good involve no agent-relativity.[3] In standard deontological theories, agent-relativity enters at the level of agent-centred constraints and prerogatives, but these are precisely the parts of such a theory that are *not* concerned with promoting the good. Thus Murphy, for example, writes that "Of all the principles that might find their way into a pluralistic moral theory, only principles of beneficence take an agent-neutral form, in the sense that they give us all the same aim" (Murphy 2000: 75).

---

[1] As an application of the principle of charity, we take it that Singer and MacAskill implicitly make this assumption, in the works cited.

[2] For discussion of the implications of standard (i.e. non-"difference making") risk and/or ambiguity aversion for these decisions, see (Buchak 2022; Pettigrew 2022; Greaves, Mogensen and MacAskill MS).

[3] Insofar, of course, as the theory of the good in question is itself agent-neutral. We will not assume that it is fully agent-neutral. For example, everything we say is compatible with the thought that benevolence properly involves giving extra weight to the interests of one's own nearest and dearest. If there is such agent-relativity, then it is obvious that benevolent agents will sometimes pull against one another. However, there will still be cases where different agents stand in the same relevant relationships to the relevant moral patients, as when two agents have the chance to rescue a perfect stranger or to benefit a common friend. In short, there will still be cases where agent-relativity is irrelevant and benevolence gives different agents, as Murphy puts it, "the same aim". All our examples involving multiple agents should all be interpreted as cases of that sort.

But a preference to make a better difference already involves agent-relativity, at least conceptually, even against the background of a single shared theory of the good. I prefer that *I* make a better difference, while you prefer that *you* make a better difference. In various ways, to be explored in this paper, this agent-relativity can - depending on further details - lead to phenomena that seem inconsistent with the spirit of benevolence.

We will stop short of asserting that difference-making preferences are *morally inappropriate*. We do not assume consequentialism; therefore, we leave it open that considerations other than those of benevolence have moral relevance (a matter we return to in section 6). Our point is more limited: where we draw a negative conclusion about particular types of difference-making preference, our conclusion is in the first instance only that the preferences in question cannot be manifestations of *benevolence.*

Despite this limitation, the point strikes us as important. For one thing, in many situations, benevolence is the most important moral consideration, to the point where departures from true benevolence in fact *are* morally inappropriate. For another, even in situations in which considerations of benevolence are not obviously overriding from a normative point of view, many people exhibit "difference-making preferences" while in fact aiming to implement, and believing themselves to be implementing, benevolence. Relatedly, we conjecture, the observations we make in this paper will give pause to many (if not all) of those who are initially drawn to the language of "making a difference".[4]

The structure of the paper is as follows. As a preliminary point, section 2 explores the consequences of thinking of the difference one makes in terms of one's direct causal impact. Perhaps unsurprisingly, preferring to make a bigger difference in *this* sense can straightforwardly make the world worse. This motivates focussing instead on the difference an agent makes in (what we will call) the "outcome-comparison" sense of difference-making. It is much more prima facie plausible that outcome-comparison difference-making preferences might be extensionally equivalent to preferences for better states of affairs. The remainder of the paper explores the ways in which this more plausible claim also fails. We begin that task in section 3, with an observation that a preference to make a bigger outcome-comparison difference can involve a preference for a worse state of affairs even in decision situations that involve no relevant uncertainty. In this case, however, the preference in question cannot lead one to prefer *actions* that tend to make things worse, and in that sense at least the tension with benevolence is limited.

But the tension between difference-making preferences and benevolence can be much stronger once uncertainty is also in play. In such a case, *one* type of difference-making preference is represented by maximisation of the expected value of the difference one makes. Difference-making preferences of that type introduce no new issues beyond those we already discuss in uncertainty-free contexts. However, in the context of uncertainty additional types of difference-making preferences are also possible (and, we believe,

[4] Some readers may balk at our invocation of something so abstract as 'the good', preferring to think of benevolence in more concrete terms of, say, promoting the welfare of particular people. Everything we say could be rephrased in such terms. In those terms, the thrust of our paper is that benevolence most fundamentally involves a preference for people's being better off, and that serious potential pitfalls lurk if, as is common, one frames would-be benevolent thinking in terms of "benefitting people" (making a difference to how well off people are).

common). Sections 4 and 5 discuss, respectively, aversion to risk and aversion to ambiguity in the difference one makes. In both these cases - especially the former - the preferences in question lead to choice dispositions that seem clearly in tension with genuine benevolence. Preferences of both types can, in a sense we will define, defeat the cause of goodness in collective action contexts. Additionally, difference-making risk averse preferences violate a condition of stochastic dominance with respect to goodness.

Section 6 discusses the issue of whether the preference patterns we have discussed, even if not purely benevolent, might nonetheless be rationalised, in terms of a desire that one's life be more meaningful. Section 7 concludes.

2 "Direct-causal" vs "outcome-comparison" difference-making

Let us start by setting aside a type of "preference to make a difference" that is rather obviously not a matter of true benevolence. Consider

> ***Example 1 (Who jumps in?).*** Abdullah and Jacinta are each passing a pond, when they see a child in danger of drowning. Given the child's situation, there is only space for one rescuer. Abdullah is better placed to conduct an efficient rescue. But Jacinta wants to do good, and consequently is keen that she rather than Abdullah gets to be the rescuer. Accordingly, she jumps in quickly, before Abdullah has a chance.

Clearly, Jacinta's preference is not an expression of pure benevolence.[5] Jacinta prefers states of affairs in which (something like) her own *direct causal impact* is better. It is better for the child if Abdullah is the rescuer. But only if she herself jumps in can Jacinta count the life saved as part of her direct causal impact.

The general lesson then seems to be that insofar as pure benevolence is a matter of preferring to make a bigger difference at all, it must be a matter of preferring to make a bigger difference in an outcome-comparison sense (which we explain next), not in the sense of direct causal impact.

The difference one makes in the outcome-comparison sense by performing some action A is a matter of how much better the outcome is if one performs A, compared to if one does nothing.[6] In this outcome-comparison sense, in our example, the difference Jacinta makes

---

[5] Similarly, Murphy writes: "[W]hereas a constraint against killing prohibits me from killing an innocent person even if this will prevent others carrying out several killings, a principle of beneficence tells me *not* to try to benefit someone myself in a situation where this will prevent others conveying greater benefits. In the former case my aim is that I not kill, not that killings be minimized. In the latter case we all have the one aim that people be benefited as much as possible, not individual aims that each of us benefit [sic] people" (2000: 75).

[6] If there is no sufficiently coherent notion of "doing nothing", or if such a notion cannot carry normative weight, this of course grounds a negative view of any difference-making preferences that essentially rely on such a baseline (such as the "difference-making risk averse" and "difference-making ambiguity averse" preferences we discuss later in the paper). In this paper, we will grant a "do nothing" baseline for the sake of argument, and argue that even if that is granted, there are strong reasons to think that the difference-making preferences that require such a baseline cannot be expressive of benevolence.
For discussion of the subtleties involved in cashing out "doing nothing" and related notions, see e.g. (Foot 1967; Bennett 1995; Woollard 2015; Woollard and Howard-Snyder 2021).

by jumping in is worse than the difference she would make by allowing Abdullah to jump in (regardless of which option is designated as "doing nothing"). Thus a preference to make a better difference in the outcome-comparison sense rationalises a strict preference, on Jacinta's part, for *not* jumping in.[7] So, in this example, there is no tension between outcome-comparison difference-making preferences and true benevolence.

We ourselves draw a further conclusion, though it is inessential to the main line of argument of our paper. In general, something like a notion of direct causal impact might have moral relevance. (For example, it seems closely connected to "integrity" in Williams' sense (1973), and to the notion of "degree of agential involvement" discussed by Wedgwood (2009).) In the particular example of *Who jumps in?*, however, benevolence should be the main concern. Because of this, Jacinta's preference for jumping in is not only non-benevolent, but is (further) morally inappropriate.

3 Preferences for higher outcome-comparison impact

Jacinta's moral mistake is to prefer that she has a higher direct causal impact, even when her having a higher impact in this sense does not make the outcome better. If she instead preferred to have a higher *outcome-comparison* impact, as noted above, she would have more benevolent choice dispositions in the case above.

In other cases, though, at least some version of preferring that one has a higher outcome-comparison impact also seems contrary to benevolence. Consider

> ***Example 2 (Mild or severe disaster)***. Jon has heard of a landslide in a neighbouring town. Some of the townspeople are trapped, and Jon has the opportunity to join the rescue effort - an effort that is close to costless for Jon. If the disaster is severe, then Jon's helping would lead to an additional 6 people being saved (compared to the case in which Jon does not help), with 14 still dying. Alternatively, if the disaster is relatively mild, Jon's helping would only save 2 additional people, with 1 still dying. Jon prefers that such a disaster be severe, because that way he gets to save more lives.

Clearly, whether the disaster is mild or severe, Jon should prefer that he help. This preference is generated by any remotely plausible version of benevolence, including difference-making preferences.

But as described above, Jon also has nontrivial preferences among the *decision situations he might face:* here, the "mild disaster" and "severe disaster" decision situations (where, in

---

It would also be possible to define a purely comparative notion of outcome-comparison difference made – the difference one makes by performing A *rather than B*. The resulting difference-making preferences would avoid the particular problems we discuss in this paper. But this doesn't seem to be the notion of difference-making that real people have in mind when they speak of a desire to "make a difference". In the context of the "difference-making risk aversion" that we discuss in section 4, it would also lead to cyclical pairwise comparisons, though we lack the space here to explore the details.

[7] The importance of focussing on the outcome-comparison rather than the causal sense of difference-making for genuinely altruistic career choice is, on a natural reading, the central point of MacAskill (2014).

each situation, the choice is whether or not to help). And Jon has a higher outcome-comparison impact - 6 additional people saved rather than only 2 - if the disaster is severe than if it is mild. If Jon always prefers worlds in which he has a higher outcome-comparison impact, he therefore prefers a world in which the disaster is severe (and he helps) over one in which the disaster is mild (and he helps), as reported in the description of the case. It seems, though, that such a preference could not be a matter of pure benevolence. As in the case of Jacinta, it seems to stem from some notion of personal moral glory, rather than from a concern with the state of the world.[8] In our own view, though again this is inessential to the main line of argument of the paper, this departure from benevolence in the example at hand is morally inappropriate.

It is instructive to compare difference-making preferences in the context of altruistic action with their analog in the context of prudential action. Suppose that instead of a landslide killing some number of strangers, Jon hears of a disease that will cut his life short. If the disease is severe, it threatens to cut 20 years off Jon's life, though he could claw back 6 of those years by engaging with a rigorous program of therapy. Alternatively, if the disease is relatively mild, it will cut 3 years off Jon's life if he does nothing, and only one year if he engages with the therapy. While Jon might gain some sliver of satisfaction from having actively preserved more years of his life, it is very clear in this case that from the prudential point of view, minimising the number of life-years lost is by far the more important thing. It would be crazy for Jon to hope for a severe illness, at the cost of dying 13 years earlier, just so that he could do more years-saving. And there seems no reason to regard the benevolent and prudential contexts as disanalogous in the relevant respects.[9]

It might nonetheless be too quick to conclude from Jon's case that difference-making preferences *per se* must be contrary to benevolence. The example depends essentially on the stipulation that Jon has difference-making preferences comparing different choice scenarios. But one could have difference-making preferences within each choice scenario without having difference-making preferences (or any preferences) between choice scenarios. Or perhaps, in a situation like Jon's, a fully benevolent agent could have a preference for the *act* of helping in a severe disaster while nonetheless having a dispreference for the *outcome*, compared to the analogous objects with a mild disaster. After all, the act really is more morally notable in the case of severe disaster, even if the outcome is horrific.

Relatedly: One might be reassured by the fact that (anyway) only Jon's preference for helping over not helping, and not any preference he might have that the disaster be severe rather than mild, is decision-relevant. In Example 2, the matter of whether the disaster is severe or mild is not under Jon's control. Furthermore, this saving grace is essential to the example. If the severity of the disaster *were* under Jon's control, then its consequences

[8] Somewhat similarly, Brian Duncan writes that an "impact philanthropist" - one who is motivated by the desire to make a difference, in contrast to either a desire for good outcomes ("public goods philanthropy") or a desire to give ("warm glow philanthropy") - "may actually benefit from need - an impact philanthropist cannot enjoy saving children if the children save themselves." (Duncan 2004: 2161)

[9] The analogy between benevolent and prudential difference-making, explored in this paragraph in the context of uncertainty-free outcome-comparison difference-making preferences, applies equally to the other types of difference-making preferences that we discuss elsewhere in this paper.

would count as part of Jon's outcome-comparison impact, so that difference-making preferences would no longer prefer a severe to a mild disaster. Perhaps it is only if the resulting choice dispositions lead to making the world worse that we can conclude that the underlying pattern of preferences is in conflict with benevolence.

We will not explore in detail whether or not any of these considerations provide a route for reconciling difference-making preferences with benevolence in the context of Example 2. Instead, we will turn to decision problems that essentially involve uncertainty.

4 Difference-making risk aversion

In this section and the next, we will argue that in the context of uncertainty, certain types of "difference-making preferences" lead even to *choice dispositions* that are contrary to benevolence. These particular types of difference-making preference are the main target of our paper.

It is important not to overstate the point we will be making. It is not that thinking in terms of "making a difference" itself *necessarily commits one* to the additional pitfalls that we will discuss, once uncertainty is in play. For example, an agent who simply maximises the expectation value of difference made to goodness (or, more generally: the difference one makes to some strictly increasing function of goodness) will avoid the pitfalls in question, and will at most face the issues discussed in sections 2 and 3. But we do think that the preference patterns we will discuss here are ones that are very natural once one combines ideas of difference-making with issues of uncertainty (specifically, with risk and/or ambiguity aversion). Also, anecdotally, these preference patterns appear to us to be fairly widespread in would-be benevolent thinking. So we do not think the points we will make attack a straw man.

We will discuss "difference-making risk aversion" and "difference-making ambiguity aversion" in (respectively) sections 4 and 5.

4.1 Difference-making risk aversion

Let us return to the cases of "Diversification" and "Cause prioritisation" that we mentioned briefly in Section 1. We will now flesh out those cases in more detail:

> ***Example 3 (Diversification).*** Mustafa has $1000 to spend philanthropically. He has narrowed his shortlist down to two charities he might support: the Against Malaria Foundation (AMF) and the Schistosomiasis Control Initiative (SCI). Both charities focus on developing-world health, but otherwise have very different missions. Correspondingly, there are different arguments for regarding the charity in question as serious candidates for "cost-effective at promoting the good". AMF distributes insecticide-treated bednets in malarial regions; the case for this intervention being highly cost-effective is primarily that it saves many lives of children under 5 per unit cost. SCI implements deworming programs; the case for this intervention being highly cost-effective is primarily that it significantly increases school attendance per unit cost. Mustafa regards each of these interventions as highly cost-effective *in*

expectation, and AMF on balance slightly more so than SCI. However, he is also highly uncertain: in fact, for each of the two interventions, he has credence 50% that the intervention in question does no net good at all in the long run. Mustafa has to decide whether to donate the whole of his $1000 to AMF, the whole to SCI, or instead to split the money between the two charities.

Example 4 (Cause prioritisation). Evie has $1000 to spend philanthropically. She is convinced that prevention of premature human extinction is extraordinarily valuable, so that the most cost-effective interventions in expected value terms are those that focus on reducing extinction risk. However, Evie also recognises that spending on extinction risk mitigation has a very high chance (very close to 1) of doing next to no good: the case for spending in this way rests primarily on the *extremely* small chance (perhaps one in a trillion trillion) that this intervention would turn out to make the difference between humanity going prematurely extinct vs. not. Alternatively, Evie could spend her $1000 instead on a "near-termist" global health intervention that has a much higher probability of doing a much more modest amount of good (for instance, via AMF or SCI).

These decision scenarios require the agent to take a stand on what "trying to promote the good" amounts to in the presence of relevant uncertainty. Standard arguments suggest that *if* Mustafa and Evie maximise *expected objective value*, then Mustafa will prefer not to diversify, and Evie will prefer to fund extinction risk mitigation (Snowden 2019: 69-71, Greaves and MacAskill 2021). At least anecdotally, however, many people facing relevantly similar real-life decision scenarios have the opposite preferences in both cases. Pressed on why that is, many are tempted to appeal to the *riskiness* of not diversifying, and of supporting extinction risk mitigation. If Mustafa fails to diversify, he faces a greater risk that his philanthropic spending does no good (50%, compared to perhaps only 25% in a diversified portfolio). Evie's case is similar but more extreme: If Evie supports extinction risk mitigation, it is almost certain that her spending will do next to no good.

This suggests that there might be a rationalisation of Mustafa's and Evie's preferences in terms of risk aversion. If an agent is risk averse with respect to some quantity X, she strictly prefers a (degenerate) gamble that delivers some particular value x* for X with certainty to a gamble that delivers an *expected* X-value of x*, but that includes nontrivial uncertainty. For example, an agent who is risk averse with respect to money prefers to receive $100 for sure than to take a gamble that delivers $50 or $150 with equal chances. More generally, a risk-averse agent strictly prefers gamble g_1 to gamble g_2 if g_2 is a mean-preserving spread of g_1.[10]

If Mustafa fails to diversify, then the spread of possible amounts of good he might generate is wider than if he diversifies; similarly (but more extremely) for Evie, if she supports extinction risk mitigation. The case in favour of diversification (resp., for prioritising global health interventions) seems precisely analogous to the standard case for diversifying one's financial investments (resp., inclining against high-stakes gambles), in a prudential context.

[10] That is, if g_1 is second-order stochastically dominant over g_2 (Hadar and Russell 1969, p.27).

For both of these lines of suggested reasoning, however, it is essential that Mustafa and Evie are averse to risk *with respect to the difference they make* ('amounts of good they might generate'). Call this *difference-making risk aversion* (DMRA). No similar case in favour of diversification and/or short-termism goes through if we instead assume risk aversion *with respect to the overall goodness of the world* (what we might call axiological risk aversion, or ARA). Indeed, one might well expect that risk aversion in the latter sense would strengthen the case for mitigating catastrophic risks. In general, this can go either way (Greaves, Mogensen and MacAskill MS). We return to this point below.[11]

DMRA preferences, however, have features that are inconsistent with true benevolence. They lead to a violation of stochastic dominance with respect to goodness, and they lead to a problematic type of collective defeat in collective action contexts. The remainder of Section 4 explores these features.

4.2. Individual action: Stochastic dominance with respect to goodness

In the absence of uncertainty, there is a clear criterion for what it takes for a set of preferences to track the betterness relation: it should be impossible, under the preference pattern in question, that A is preferred to B, yet B leads to a better outcome than A. This is the condition that is violated by Jon's difference-making preferences in *Mild or severe disaster,* albeit not by his preferences between options within a choice situation.

In the presence of uncertainty, things are more complicated. We have a set S of states of nature, such that the outcome of any given action depends on which element of S is actual, but the agent is uncertain which element of S is actual. Actions, in this context, are often modelled as functions from S to O, where O is the set of possible outcomes. The theory of the good gives us (at least) a betterness ordering $\succeq$ on O. The agent faces a choice set C, containing all actions that are available to the agent in the decision situation in question.

In discussing the notion of benevolence, we want to focus on which actions it is reasonable for the agent to prefer for choice purposes, given her desire to promote the good, but also given the fact that this uncertainty will not be resolved before she makes her choice. For this focus, it would not do simply to note that (say) if $A(s_1) > B(s_1)$ but $A(s_2) > B(s_2)$, then the agent prefers A to B *conditional on state* s_1, but prefers B to A *conditional on state* s_2. We also need a notion of unconditional act-preference. What constraints must the latter satisfy, if it is to track betterness, and thereby to be a plausible candidate for an expression of pure benevolence?

[11] A decision theory according to which agents (in some sense to be spelled out) "ignore tiny probabilities" might also rationalise Evie's reluctance to devote her resources to extinction risk mitigation. But note that the tiny probability in Evie's case is the probability that she makes a difference, not the baseline probability of extinction. So again this line of thought depends on applying the decision-theoretic idea in question (here, "ignore tiny probabilities") specifically to considerations of the difference the agent makes, rather than directly to considerations of outcome goodness. It also leads to problems very similar to those we discuss for risk aversion in this section (Kosonen 2022, chapter 6).

A minimal condition is that it should be impossible, under the preference pattern in question, that A is preferred to B while B leads *with certainty* to a better outcome than A. That is, the preference ordering $\succeq_p$ must *respect statewise dominance*, a condition we now define:

> **Statewise dominance with respect to goodness (definition).** For $A, B \in C$, B *statewise dominates* A iff:
>
> - For all states $s \in S$, $B(s) \geq A(s)$; and
> - For at least one state $s \in S$ that has positive probability, $B(s) \succ A(s)$.
>
> **Respect for statewise dominance with respect to goodness (definition).** A preference ordering $\succeq_p$ on C *respects statewise dominance* iff for all $A, B \in C$, if B statewise dominates A, then B $\succ_p$ A.

We take it that respect for statewise dominance with respect to goodness is a necessary condition for preferences to count as tracking betterness. However, this condition is so weak that it often does little to discriminate between otherwise plausible preference patterns. In particular, in the context of individual action it is satisfied by DMRA preferences.

It is also extremely plausible, however, that for preferences to count as tracking betterness, and therefore for them to be a plausible candidate for pure benevolence, they should satisfy at least one further condition: respect for *stochastic* dominance with respect to goodness. Let us explain this condition using the following example.

> ***Example 7 (Swap and sweeten)*.** A fair coin is to be flipped. The agent's choice set is {Do nothing, A, B, C}, with possible outcome-goodness as follows:

Outcome goodness	**Heads**	**Tails**
Do nothing	10	0
A	20	10
B	10	$20 + x$
C	10	20

We claim that any benevolent agent should prefer B to A. There are several different, although essentially equivalent, ways to explain the case for this claim. First, note that B statewise dominates C, so a benevolent agent should strictly prefer B to C. But when it comes to the goodness of the world, A and C are clearly equivalent: for both options, there is a ½ probability of an outcome with goodness 10, and a ½ probability of an outcome with goodness 20. A and C differ only over the *correlation* between states of the world and outcomes, and (the point of focussing on stochastic rather than only statewise dominance) this correlation does not itself matter. So, a benevolent agent should be indifferent between A and C, and thus should prefer B to A. Second, we could note more directly that, when it comes to outcome-goodness, B differs from A in only one important way: instead of a ½ probability of getting 20, B involves a ½ probability of getting 20+x. In short, compared to A, B shifts probability from a worse to a better outcome; this is clearly an improvement. Finally,

we could note that B gives at least as great a probability as A, and in some cases a *greater* probability, of meeting or exceeding any given benchmark for goodness: they both give a ½ probability that the world will be at least as good as 10 and a ½ probability that the world will be at least as good as than 20, but only B gives any positive probability that the world will be at least as good as 20+x.

All this is to say that B *stochastically dominates* A, in the following sense:

> **Stochastic dominance with respect to goodness (definition).** For $A, B \in C$, B *stochastically dominates* A *with respect to goodness* iff:
>
> - For all outcomes o in O, the probability that the outcome is at least as good as o is at least as high on B as it is on A; and
> - For some o in O, the probability that the outcome is at least as good as o is strictly higher on B than it is on A.

And, we propose, a benevolent agent's preferences should respect stochastic dominance:

> **Respect for stochastic dominance with respect to goodness (definition).** A preference ordering $\succeq_p$on C *respects stochastic dominance with respect to goodness* iff for all $A, B \in C$, if B stochastically dominates A, then $B \succ_p A$.

The problem is that DMRA preferences do not respect stochastic dominance with respect to goodness. To see this, reconsider options A and B in *Swap and sweeten.* The difference to goodness made by the agent, for each relevant combination of action and state of nature, is as follows:

Difference made to goodness	Heads	Tails
A	10	10
B	0	$20 + x$

For any DMRA agent, there is some (sufficiently small) value of x such that the agent strictly prefers A to B.[12] But for all positive values of x, B stochastically dominates A with respect to

[12] This is true for any type of risk aversion satisfying the general definition above, provided that preferences are continuous in goodness. If the agent always strictly disprefers a mean-preserving spread in goodness, then she strictly prefers A to B when x=0 (that is, despite our argument above that A and C are transparently equivalent from the point of view of goodness, she strictly prefers A to C). Therefore, if preferences are continuous, there is also some positive value x>0 such that the agent still strictly prefers A to B.
To make things more concrete, we would need to consider a particular model of risk aversion. By way of illustrative example, let us consider risk aversion in the sense that is compatible with expected utility theory. On this account, to be risk averse with respect to difference made to goodness is to have a utility function that is a strictly concave function of difference made to goodness. Suppose the agent's utility, in any given outcome, is equal to the square root of the difference she herself makes in that outcome. Then, if she maximises expected utility, she prefers A to B provided x<2.5.
This is not the only available model of risk aversion. See O'Donoghue and Somerville (2018) for a recent survey of such models, including the rank-dependent approach that has been championed in the philosophical literature by Buchak (2013).

goodness. Therefore DMRA preferences fail to respect stochastic dominance with respect to goodness.

We submit that any pattern of preferences that is plausibly seen as purely benevolent must respect stochastic dominance with respect to goodness. If B stochastically dominates A in that sense, then B is superior to A as far as the goodness of the world is concerned; so any pattern of preferences that is purely motivated by benevolent concern for the state of the world must prefer B to A.[13] Therefore DMRA preferences are not purely benevolent.

4.3 Collective action: Collective defeat of goodness

The examples we have considered so far each focus on the actions of a single agent. While other agents are involved in some of the examples (for instance, Abdullah as well as Jacinta in *Who jumps in*?), the choices made by those other agents, in those examples, have a status similar to that of "nature's choice" between states of nature. In particular, it is of no relevance to the above discussions whether or not any other agent, besides the one we focus on, also has difference-making preferences.

Additional tensions between DMRA preferences and the notion of tracking goodness arise when we consider the gambles that are generated at a collective level, considering the actions of more than one such "difference-making" agent together. Consider:

> ***Example 8 (Anticorrelated altruistic risks):*** Two agents, Rio and Renzo, face a choice between altruistic interventions *Safe* and *Risky*. They both take goodness to be linear in lives saved. Each instance of *Safe* saves two lives for sure, though it also brings about a harm that is equivalent to the loss of x lives, for some small quantity x>0. Each instance of *Risky* saves either 0 lives or 4 lives. The two instances of *Risky* are anticorrelated, so that with certainty, if both agents choose *Risky*, one saves 0 lives while the other saves 4 lives.

In terms of goodness, the outcome matrix for this collective action problem is as follows:

Outcome goodness	**Rio does nothing**	**Rio chooses *Safe***	**Rio chooses *Risky***
Renzo does nothing	0	$2-x$	0 or 4, each with probability ½
Renzo chooses *Safe*	$2-x$	$44-2x$	$2-x$ or $6-x$, each with probability ½
Renzo chooses *Risky*	0 or 4, each with probability ½	$2-x$ or $6-x$, each with probability ½	4

Let us suppose that Rio and Renzo have identical DMRA preferences. Will they prefer to perform an instance of *Safe*, or an instance of *Risky*? Well: at the individual level, for any

[13] There is of course no tension between stochastic dominance and risk aversion in general. Relatedly, DMRA preferences do respect a different type of stochastic dominance, viz. stochastic dominance with respect to the difference the agent makes to goodness (as opposed to: with respect to goodness itself). But this fact does nothing to undermine our point.

positive value of x, an instance of *Risky* leads to a greater increase in goodness *in expectation* (the equivalent of 2 lives saved, rather than $2 - x$ for each instance of *Safe*). But *Risky* also has a higher spread of possible differences it might make. Because of this, for any DMRA agent, there is some $x > 0$ such that the agent prefers *Safe* to *Risky*.

So (with x specified appropriately), Rio and Renzo will both choose *Safe*. The predictable result is that while 4 lives are saved, in addition $2x$ units of harm ensue (with $x > 0$). Equally predictably, if they had both instead chosen *Risky* then 4 lives would have been saved *without* the additional harm - an outcome that both agents recognise to be better.

Let an *option profile* be an assignment of options to agents. The upshot of Example 8 is that when the outcome is jointly determined by the choices of more than one agent, agents acting in accordance with DMRA preferences can lead to an option profile that is statewise dominated by some alternative option profile that was collectively available to them. We will say that DMRA preferences lead to "collective defeat of goodness", a notion that we will later make this notion more precise.[14] Insofar as the agents aim to be engaging in benevolence, this should trouble them.

4.4 Varieties of defeat under pure benevolence

We said that susceptibility to collective defeat of goodness "should trouble" an agent intended to engage in benevolence. This much seems right. But does the "troublingness" in question amount specifically to a *reason to think that DMRA preferences cannot be purely benevolent*? Or is this just another example of the mundane fact that there are situations in which even pure benevolence fails to achieve its aim? We will argue it is the former.

To examine this issue, let us first clarify the types of "defeat" that can occur even under ordinary benevolent preferences - that is, for our purposes, when the preferences of all agents simply track betterness, according to a shared conception of the good (and there is, therefore, no special role for a notion of "making a difference").

First, let us assume that all agents are ordinarily benevolent *and risk neutral* with respect to a shared conception of the good. The following three examples illustrate that even then, suboptimal outcomes can ensue.

> ***Example 9 (false beliefs about the actions of other agents)****. Douglas and Gustav want to meet for lunch. They are both indifferent between curry and pizza, and prefer meeting to not meeting. They cannot communicate. Douglas thinks that Gustav will*

[14] Our point concerns collective defeat *of goodness*, not collective *self*-defeat. The latter occurs when (perhaps subject to further conditions) agents each acting in accordance with their own preferences jointly bring about an outcome that *both agents disprefer*. It is well known that agent-relative preferences can lead to collective self-defeat. This is illustrated by, for example, the Prisoners' Dilemma.
Collective self-defeat does not occur in *Anticorrelated altruistic risks*. Although any possible outcome from one agent choosing Risky is better than the sure outcome if both agents choose Safe, for any such outcome, either Rio or Renzo (whichever of them turned out to save zero lives) disprefers it. As one would expect given the agent-relativity, DMRA preferences can also lead to collective self-defeat (in other cases), but this fact is not interesting for present purposes.

go for curry, so Douglas goes for curry. Gustav thinks that Douglas will go for pizza, so Gustav goes for pizza. They fail to meet.

In Example 9, there is no relevant uncertainty about states of nature (as opposed to: about the actions of other agents in the decision problem). Both agents know that the outcome of (Douglas to curry, Gustav to pizza) is worse than that of each of the two option profiles in which both agents go to the same place. Yet, (Douglas to curry, Gustav to pizza) is the option profile that occurs. The lesson of Example 9 is that *false beliefs about the actions of other agents* can lead to a form of defeat of goodness, even under ordinary benevolence.

Next, consider

Example 10 (true beliefs that other agents will play their part in a suboptimal Nash equilibrium)*. Malin and Jill want to meet for lunch. They both prefer curry to pizza, but more strongly than that, they prefer meeting to not meeting. They cannot communicate. Malin thinks that Jill will go for pizza, so Malin goes for pizza, and similarly vice versa. They meet, but would both have preferred to meet at the other location.*

Like Example 9, Example 10 involves no relevant uncertainty. Unlike Example 9, in Example 10 each agent's belief about the other agent's action is accurate. But in Example 10 there is a *suboptimal Nash equilibrium,* namely the option profile (pizza, pizza).[15] Because it is a Nash equilibrium, if each agent believes that the other will play her part in that equilibrium, the agent's best response is to follow suit. But because it is a *suboptimal* one, in doing so the agents collectively bring about an outcome that is worse than the best available.

Intuitively, Malin and Jill could have coordinated on the optimal Nash equilibrium if they had been able to communicate. In Example 10 as specified, each agent expects the other to go for pizza, despite the common knowledge that both agents prefer curry. The reasons for this expectation are unspecified, but it is not hard to fill in the details: something in the background is making the pizza restaurant function as a Schelling point for Malin and Jill (for example, perhaps the pizza place is the one regularly frequented by their department seminar).[16] But a simple phone conversation could have fixed that. No commitment mechanism is needed, since (curry, curry) is also a Nash equilibrium: the role of the phone conversation is only to shift the Schelling point.

Thirdly, consider

Example 11 (Different credences about states of nature). Thomas and Lél want to meet for lunch. They both prefer curry to pizza, but they have complicated and

[15] An option profile A is a *Nash equilibrium* if, conditional on the assumption that all other agents play their respective parts in A, each agent prefers also to play her part in A. By a "suboptimal" Nash equilibrium, we mean one that is Pareto dominated by another Nash equilibrium in the same decision problem.

[16] In the context of games of cooperation, where there are multiple Nash equilibria, a Schelling point is a Nash equilibrium that is somehow salient to all agents, such that agents are able to coordinate on that particular equilibrium even in the absence of communication by all playing their respective parts in the Schelling point. (Schelling 1960:57)

confusing social relationships with the owners of the curry and pizza restaurants. Thomas thinks that because of these relationships, great happiness would ensue if Thomas went to pizza alone and great suffering would ensue if Lél did, while Lél thinks the reverse. Because of this, they have different ex ante evaluations of the two option profiles in which they fail to meet. Their expected utilities for the four option profiles have the structure of a prisoner's dilemma:

(Thomas's assessment, Lél's assessment)	**Thomas to curry**	**Thomas to pizza**
Lél to curry	(good, good)	(best, worst)
Lél to pizza	(worst, best)	(OK, OK)

In terms of their ex ante assessments, going to pizza is a dominant option for Thomas, and similarly for Lél. They both go for pizza.

The point of Example 11 is that the game-theoretic structure of a multi-agent decision problem is determined by the *ex ante* assessments agents make of *option profiles,* not directly by their *ex post* evaluations of *outcomes*. Differing credence distributions over states of nature can make it the case that the agents have arbitrarily different ex ante assessments for the available option profiles, even if they share a common evaluation of outcomes (for example, if they are ordinarily benevolent and risk neutral with respect to a shared theory of the good). In Example 11, in terms of *ex ante* assessments, the decision problem has the structure of a prisoners' dilemma, and a suboptimal result ensues.

A fourth type of example is possible if we drop the restriction that the agents must be *risk neutral* with respect to a shared cardinal theory of the good, and require only that they have a common *ordinal ranking* of outcomes:

Example 12 (Differing degrees of risk aversion). A fair coin is to be flipped. There are two agents, Hong and Liu. The goodness of the outcome attendant on either resolution of the coin toss depends on the choices of two agents taken together, as per the following outcome matrix:

Lives saved	**Liu does A**	**Liu does B**
Hong does A	50 if Heads, 150 if Tails	y if Heads, $200 + y$ if Tails
Hong does B	$100 - 2x$	$50 - x$ if Heads, $150 - x$ if Tails

Hong is risk averse with respect to outcome goodness, while Liu is risk neutral.

In Example 12, for any positive degree of risk aversion Hong might have, there are some positive values x, y such that Hong prefers the option profile in which she does B, regardless of what Liu does. And since Liu is risk neutral, for any positive values x, y, Liu prefers the option profile in which she does B, regardless of what Hong does. That is, for some positive x and y, B is a dominant option for both agents. Yet for any positive x, the outcome of both agents doing A is certainly better than that of both doing B.

The lesson of Example 12 is that even if agents are ordinarily benevolent with respect to a shared theory of the good, if they have differing degrees of risk aversion with respect to that goodness scale, again (as in Example 11) their *ex ante* assessments of option profiles can have the structure of a prisoner's dilemma, so that rational choice leads to a suboptimal outcome.

4.5 Taking stock

Examples 9-12 illustrate ways in which some form of defeat of goodness can occur even when agents are ordinarily benevolent with respect to a shared theory of the good, if they have unfortunate beliefs about each others' actions, differing credence distributions over states of nature, or differing degrees of risk aversion.

The following definition, however, excludes examples 9-11. Let a *preference profile* be an assignment of preferences to agents. We then define:

> **Strong collective defeat of goodness (definition).** Say that a preference profile P exhibits *strong collective defeat of goodness* iff for some collective decision problem involving identical credence distributions over states of nature, every option profile that is a Nash equilibrium according to P is statewise dominated in terms of goodness by some other option profile.

Strong collective defeat of goodness is provably impossible for preference profiles in which all agents are expected utility maximisers with a common cardinal utility function, and therefore for preference profiles in which all agents are ordinarily benevolent and risk neutral with respect to a shared cardinal theory of the good.[17]

This suggests the following argument:

Collective Defeat Argument:

1. Truly benevolent preferences cannot exhibit strong collective defeat of goodness.
2. DMRA preferences exhibit strong collective defeat of goodness.

   Therefore,
3. DMRA preferences are not truly benevolent.

One might worry that the above definition of strong collective defeat is obviously cooked up to exclude by fiat the known ways in which even ordinary benevolence can lead to something like collective defeat, and that the first premise of this argument is therefore question-begging. On the other hand, in Examples 9-11, there is an obvious 'locus of blame' for the failure to attain an optimal outcome. In Example 9, intuitively, things go wrong not because the agents' preferences are insufficiently benevolent, but because they

[17] Proof: If the agents have the same utility function on outcomes and also the same credence distribution on states of nature, then they assign the same *expected* utility as one another to every option profile. Let $A = (a_1, \dots, a_n)$ be an option profile that has maximum attainable expected utility according to the agents' shared credence distribution. Let $i \in \{1, \dots, n\}$ be an arbitrary agent. Suppose that all other agents apart from i will in fact play their respective parts in A, and that agent i has credence 1 that they will do so. Then, since A has maximal expected utility according to i, the best response set for agent i contains a_i. So $A = (a_1, \dots, a_n)$ is a Nash equilibrium.

(blamelessly) have false beliefs about what each other will do. In Example 10, things go wrong because, unfortunately, some force makes a suboptimal Nash equilibrium into a Schelling point, and the agents lack the communicative resources to overturn this. In Example 11, things go wrong because the agents have different views of the empirical structure of the world, and this divergence in descriptive beliefs forms a barrier to effective collective action. But in *Anticorrelated altruistic risks*, there is nowhere to lay the blame except at the door of the preferences themselves: intuitively, in this example the suboptimality is a result of the agents being overly concerned with their own place in the world, and insufficiently concerned with the world itself. From this point of view, the exclusions in our definition of strong collective defeat of goodness are entirely justified: they exclude non-preference-based sources of unfortunateness that can block the attainment of optimality given *any* type of preferences, so that any remaining phenomena of "defeat of goodness" can be blamed on the preferences themselves.

Example 12 *is* a case of strong collective defeat of goodness as defined above. There are two reasonable views about this case that are compatible with our main argument. Defeat arises here from the *difference* between the two agents' degrees of risk aversion (with respect to their shared cardinal theory of the good). In line with the Collective Defeat Argument, one could take this to show that at least one of the agents does not have truly benevolent preferences (perhaps benevolence turns out to require a particular degree of risk aversion, e.g., risk neutrality). Alternatively, one could amend our notion of defeat to exclude this type of example as well. After all, the difficulty for Liu and Hong arises from their failure to coordinate their degrees of risk aversion, much as the difficulty for Thomas and Lél in Example 11 arises from their failure to coordinate their credences. In contrast, in *Anticorrelated altruistic risks*, it is the very fact that the agents are averse to risk in the difference they themselves make that leads to the suboptimality. There does not seem to be any sense in which Liu and Hong could coordinate without fundamentally altering the character of their preferences.[18]

We do not claim that the Collective Defeat Argument, in either original or revised form, is entirely watertight. There is logical space (at least) to hold that there is simply no tension between benevolence and susceptibility to collective defeat, even when conditions are such as to make collective defeat eminently avoidable. On the other hand, this position does seem unattractive. Whatever one thinks of the case for virtue consequentialism in general (Driver 2001; Calder 2007), it seems less open to question that *benevolence* earns its place among the virtues via the fact that agents' possession of this trait conduces to the good under normal conditions. It also seems that this should hold when the issue is what results from *multiple* agents possessing the trait in question: there is nothing essentially individualistic about the consequentialist criterion for some trait to count as a virtue. So our own view is that the (original or revised) Collective Defeat Argument is sound: susceptibility to strong collective defeat of goodness is an additional reason to regard DMRA preferences as non-benevolent, over and above the violations of stochastic dominance noted in section 4.2.

[18] If the agents in some group are averse to risk with respect to the difference they make *together*, then strong collective defeat of goodness cannot occur *within that group*. This would be another way for Rio and Renzo to avoid choosing the suboptimal option profile (Safe, Safe) in Example 8. But one might wonder then why the relevant group should be just Rio and Renzo, rather than (say) all agents. It seems likely this line of thought will lead away from difference-making preferences and back towards ordinary benevolence, though we lack the space to explore this here.

5. Difference-making ambiguity aversion

So far, we have considered agents who deviate from expected value maximisation at most by being risk averse. In this section, we consider a different type of deviation: ambiguity aversion.

An ambiguity averse agent tends to prefer gambles in which the probabilities are more objectively constrained. To illustrate, consider:

> ***Example 13 (Unambiguous vs. ambiguous bet).*** One urn contains 50 red balls and 50 black balls. A second urn is known to contain 100 balls, each of which is either black or red, but the proportions are unknown. Somchai chooses whether to draw a ball at random from the first or the second urn. He receives $100 if the ball he draws has a specified colour; otherwise, he receives nothing.

If Somchai is ambiguity averse, he might strictly prefer to bet on the first urn, where he knows the chance of winning, regardless of which colour he is betting on. This preference seems inconsistent with expected utility theory, but is widespread (Ellsberg 1961; Trautmann and Kuilen 2015).[19]

As in the case of risk aversion, it is natural to think that ambiguity aversion might rationalise both a preference for diversifying of one's "portfolio" of attempts to make a difference, and a preference for mitigating existential risks over relatively "near-termist" interventions (such as funding AMF or SCI). We will explicate these two points in turn.

Regarding diversification, consider:

> ***Example 14 (Making friends with power).*** Important national elections will take place in a few months. The two main parties seem symmetrically placed vis-a-vis their prospects for winning the election, but there is significant ambiguity on the matter of which party will win. Whichever party wins, Catalina will at a later date be able to improve the nation's welfare in proportion to the strength of her relationship

[19] As in the case of risk aversion, there are several models of ambiguity aversion. For concreteness, we will sketch one such model: the α-maxmin model. In this model, each agent is represented by a utility function U on the set O of outcomes, a class R of probability functions on the set S of states (the agent's representor), and a parameter $\alpha \in [0,1]$. The agent's preference ordering over gambles $a: S \rightarrow O$ is represented by the formula

$$(1-\alpha)\max_{p \in R} EU_p(a) + \alpha \min_{p \in R} EU_p(a),$$

where $EU_p(a)$ is the expected utility of the gamble a with respect to the probability function p. In this model, the parameter α is a measure of ambiguity aversion, akin to pessimism. If $\alpha = 0$, the agent prefers whichever gamble has the highest 'maximum possible' expected utility, generated by any of the probability functions in R; such an agent is ambiguity-seeking. If $\alpha = 1$, the agent prefers whichever gamble has the highest `minimum possible' expected utility; this is the most extreme form of ambiguity aversion that is possible in this model. Values of α in the range $(0.5, 1)$ correspond to less extreme forms of ambiguity aversion.
In Example 13, for instance, let us suppose that Somchai's representor R contains elements according to which the probability of drawing a black ball is anything between 40% and 60% (inclusive). Then, as is straightforward to verify by calculation, he will prefer the ambiguous to the unambiguous urn if the value of his ambiguity aversion parameter α is greater than 0.5.

with the winning party. But such a relationship can be built only during the run-up to the election.

In Example 14, the thought runs, ambiguity aversion might incline Catalina more towards doing some relationship-building with both parties, rather than directing all her efforts at just one party (even if the strength of relationship built is linear in relationship-building effort expended). For this enables her to hedge against the ambiguity inherent in the election result, and to attain a situation in which the amount by which she can improve things post-election is the same regardless of which party wins, and is thus unambiguous.

Regarding cause prioritisation, the key thought is that for carefully selected near-termist interventions, we have robust empirical guidance constraining the probabilities of the intervention's possible effects. In contrast, attempts to mitigate extinction risks cannot be guided in anything like the same way by experimentation or feedback; their effectiveness seems often to a far greater extent a matter of guesswork.[20] Therefore, the thought runs, there is significantly more ambiguity about the payoff from existential risk mitigation efforts than there is from efforts directed towards more near-term ends.

As in the arguments for diversification and for deprioritising extinction risk mitigation on grounds of "risk aversion", however, the arguments from "ambiguity aversion" in these examples implicitly presuppose that the aversion is to ambiguity in the difference one's own intervention makes (what we will call "difference-making ambiguity aversion", or DMAA), rather than to ambiguity in the overall value of the world ("axiological ambiguity aversion", or ARA). Regarding diversification: in Example 14, aversion to ambiguity in the total value of the world would tend to favour building relationships exclusively with that party, whichever it might be, whose victory would be expected to lead to a worse future in the status quo. Regarding cause prioritisation: it is already extremely ambiguous how much risk of premature extinction humanity faces, and there is no particular reason to think that interventions aimed at reducing this background risk increase, rather than decrease, ambiguity in the latter sense.[21]

This raises the question, as discussed above for risk aversion, what can be said about the credentials of difference-making ambiguity aversion (DMAA) from the point of view of benevolence.

In the case of difference-making risk aversion (DMRA), we argued

(1) That even at the individual level, DMRA preferences fail to respect stochastic dominance with respect to goodness, and in that sense fail to track the good.

---

[20] Specifically, this is generally the case for attempts to mitigate *anthropogenic* extinction risks, such as those from nuclear war, biotechnology or advanced artificial intelligence. Cost-effectiveness estimates for programs to mitigate extinction threats from asteroids are relatively strongly objectively constrained. See, e.g., Greaves and MacAskill (2021, section 4).

[21] The implications of ARA for the evaluation of existential risk mitigation are investigated in Greaves, MacAskill and Mogensen (MS).
An orthogonal objection to the argument from (even difference-making) ambiguity aversion to a preference for "near-term-motivated" interventions over existential risk mitigation is that the former, no less than (and plausibly more than) the latter, involve enormous ambiguity regarding their further future effects (Greaves (2020)). Here we set that aside.

(2) That when collective action is involved, DMRA preferences are susceptible to strong collective defeat of goodness, in a way that is inconsistent with pure benevolence.

There is no analog of the first point for DMAA. While it is a little delicate what stochastic dominance amounts to in the presence of ambiguity, on reasonable ways of making it precise, it is entirely possible for a difference-making ambiguity averse agent to always prefer a stochastically dominant option to a stochastically dominated one.[22]

However, as we will show next, the issue of collective defeat plays out very similarly for ambiguity aversion as it does for risk aversion. To see that DMAA preferences give rise to strong collective defeat of goodness, consider:

> ***Example 15 (Unambiguous vs. ambiguous lifesaving)****.* Two urns are as in Example 13*.* One ball is about to be drawn at random from each urn. Bussaba and Rochana each face the choice of which urn to bet on. Bussaba (resp. Rochana) wins her bet if a black (resp. red) ball is drawn from her chosen urn. For each winning bet on the unambiguous urn, 10 lives are saved. For each winning bet on the ambiguous urn, an amount of goodness ensues equivalent to saving $10 + x$ lives.

As in Example 13, an agent who is ambiguity neutral would tend to prefer the ambiguous urn, while an agent who is averse to ambiguity in the number of lives she herself saves will tend to prefer the unambiguous urn.[23]

Note, though, that if both agents opt for the ambiguous (resp. the unambiguous) urn, the outcome is certainly that 10+x (resp. only 10) lives are saved. Thus, like difference-making risk aversion, difference-making ambiguity aversion can lead to strong collective defeat of goodness.

Also analogous to the case of risk aversion: this route to collective defeat of objective goodness arises only for *difference-making* ambiguity aversion. It has no analog in the case of goodness-increasing ambiguity aversion.[24]

---

[22] The issue is that in the presence of ambiguity, there may not straightforwardly be any such thing as *the* probability of a given outcome on a given action. But, for example: in the alpha-maxmin model of ambiguity aversion (cf. fn. 19), if utility is given by difference made to a strictly increasing function of goodness, then even an ambiguity averse agent will always prefer one option to a second if the first stochastically dominates the second according to every probability function in the representor.

[23] For example, let us again assume the α-maxmin model of ambiguity aversion, and suppose that Bussaba and Rochana's credence that a ball drawn from the ambiguous urn would be black is represented by the interval $[0.4, 0.6]$. If (say) $x = 0.5$, then the agents prefer the unambiguous urn if $\alpha > 0.62$.

[24] In Example 14, for each agent, whether selecting the unambiguous or the ambiguous urn leads to more ambiguity in total value depends on which urn the other agent selects. The total outcome of both agents selecting the ambiguous urn is unambiguous ($10 + x$ lives are saved for sure). Because of this, if either agent selects the ambiguous urn, the other agent's best response tends to be to do likewise, if the agents are averse to ambiguity in total value but not in the difference they themself make.

6. The appeal to meaningfulness

Let us take stock. We have canvassed several types of predicament in which various versions of a desire to make a difference seem inconsistent with pure benevolence. Jacinta's desire to be the rescuer does the drowning child no good, and might even do harm (if Abdullah is a more effective rescuer). Jon's desire to save more lives can lead to a preference that worse disasters occur, which at least threatens to be inconsistent with pure benevolence (though in this case, the apparently anti-benevolent preference at least cannot lead his *actions* astray). Mustafa and Evie's aversion to risk in the difference they make can lead to violations of stochastic dominance with respect to goodness at the individual level, and to collective defeat of goodness at the collective level; similarly for "difference-making ambiguity aversion".

In at least some of these cases, however, anecdotally, intuitive agential preferences for the behaviour described remain widespread, even after the tensions with pure benevolence are understood. What should we make of this?

A natural diagnosis is that what is going on in all these cases is a sort of pseudo-benevolence, rather than the genuine article. Perhaps, for instance, what is really driving the preferences in question is a desire to render one's own life in some sense more meaningful, rather than a direct desire that the world be better. It might then not be surprising that the corresponding behaviour is in tension with benevolence, since a desire for meaningfulness is not straightforwardly a matter of benevolence.

We will explore how this pans out in the context of the various sorts of "desire to make a difference" that we have discussed above.

First, consider Jacinta's preference to be the rescuer. Assuming only that she is a minimally decent human being, Jacinta will take some satisfaction in the knowledge that the child is saved, even if it is Abdullah rather than Jacinta herself who does the saving. Nonetheless, Jacinta might well feel that her own day is made more satisfying in a morally laden sense, and (relatedly) her own life made more meaningful, if it is she herself who directly does the rescuing. Thus, Jacinta's preference is comprehensible in terms of a desire for meaningfulness.

Second, consider Jon's preference for a severe disaster. This is a case of preference to make a better difference in the outcome-comparison rather than in the direct-causal sense, yet the remarks we have just made about Jacinta seem equally well to apply here. While Jon's preference is for a worse state of affairs - one in which more innocents die rather than fewer - and in that sense is straightforwardly anti-benevolent, yet the preference is eminently comprehensible in terms of a desire Jon might have that his own life be morally significant.

Finally, let us consider the case of difference-making risk aversion (again, similar comments apply to difference-making ambiguity aversion). If, say by donating to AMF, I make the difference between hundreds of young children dying young versus surviving to full adulthood, then (the thought runs) I will have done something significant with my life. But if I devote my resources to extinction risk mitigation and, as things turn out, this makes no difference *ex post*, then in at least one clear sense all my sacrifice has been for nought.

Arguably, in this (overwhelmingly probable) case, my sacrifice of material well-being has then not brought with it any compensating increase in the meaningfulness of my life.[25] In this frame of mind, it is very natural, insofar as one is inclined towards risk- and/or ambiguity aversion, to be averse to risk or ambiguity in the difference one makes, rather than to risk or ambiguity in the value of the world largely independently of one's actions.

So, in all three cases, it is at least *possible* to rationalise "difference-making preferences" of the kinds we have discussed in terms of a desire for meaningfulness, or moral significance, in one's own life. We further hypothesise that something like the above is fairly descriptively accurate, as an account of the psychology of at least (i) many would-be value-maximising philanthropists who feel the pull of the intuitions towards diversification and towards short-termism,[26] and (ii) many who arrange that they themselves have a positive direct causal impact, and value this enormously out of proportion to the extent to which the arrangement does counterfactual good.

On reflection, however, the motivations in question seem questionable, perhaps even in terms of internal coherence. Again, we are not presupposing consequentialism. We do not say, therefore, that any preference that runs contrary to benevolence is *ipso facto* morally inappropriate, nor that no such preference can arise from a desire for meaning in one's life. But very often, agents engaged in difference-making reasoning *are motivated by the ideal of benevolence*. They are *trying* to lead lives that incorporate that virtue, and that is what drives their attempts to make a positive difference. If an agent meeting *this* description (at least) is convinced (perhaps by the arguments of this paper) that difference-making preferences of some given form run actively contrary to benevolence, it seems incoherent then for such an agent to hold that their life is made more meaningful by "engaging in the project of benevolence" by having and implementing preferences of that particular form, if the alternative is genuine benevolence.[27] It seems more meaningful to live in accordance with genuine benevolence, rather than some semi-egoistic halfway house. Relatedly, pure rather than pseudo-benevolence seems the more morally laudable trait.

It is instructive here to compare the above discussion with the philosophical discussion of Williams' thought-experiment of "Jim and the Indians" (Williams 1973). In the latter case, while most authors agree that all things considered, Jim should shoot (as recommended by benevolence), non-consequentialists generally hold that there are very strong pro tanto reasons for Jim not to shoot. The important point for us is that for the non-consequentialist, this is a case of a deep and troubling conflict between reasons of benevolence on the one hand, and moral reasons *of some other type* on the other.[28] But there is less room to motivate difference-making preferences in this way than there is to motivate Jim's reluctance

[25] For this reason, Monton (2019: 15-16) argues that it is not rational for an altruist to maximise expected value in such 'Pascalian' cases.

[26] We say *many* rather than *all* such philanthropists for two reasons. (1) There are also many who simply disbelieve the arguments for the claim that interventions to mitigate existential risk lead to higher expected objective value than global health interventions. Our discussion is rather of why one might feel a pull towards the latter *even if one agrees* with the arguments in question. (2) We do not deny that there may also be other possible explanations for the intuitions in question: for example, in the case of cause prioritisation, a sense that global health interventions are more strongly morally demanded by considerations of compassion or of justice.

[27] A similar objection is levelled against the coherence of "warm-glow philanthropy" by Elster (2011).

[28] For a list of possible non-consequentialist reasons not to shoot, see e.g. Kamm (1999:173).

to shoot, because again, unlike Jim's anti-shooting preference, difference-making preferences are often intended from the outset to be expressions of benevolence.

7. Summary and conclusions

Pure benevolence is a matter of preferring A to B iff, and then because, A is better than B. Agents motivated by ideals of benevolence, however, often tend to at least frame their discussions in terms of preferences to make a (better) *difference* to goodness. This is at least conceptually quite a different matter; unlike a preference that a better state of affairs obtains, a preference that one oneself makes a better difference is an agent-relative concern.

The ideology of difference-making is nonetheless innocuous insofar as such "difference-making preferences" are extensionally equivalent, at least as far as action-relevance is concerned, with preferences for better states of affairs. But in many of the forms that such "difference-making preferences" often take, that extensional equivalence fails.

First, preferences to make a better difference in (what we called) the "direct causal" sense clearly come apart from preferences for better states of affairs. This is because they assign importance to the agent's own causal role in a way that is not rationalised by, and can run counter to, a preference for better outcomes (for example, a preference that one try to rescue a child oneself, rather than allowing some more effective rescuer to do it).

This observation motivates defining a notion of difference made in (what we called) the "outcome-comparison" sense. But, second, even a preference to make a better difference in this less causally loaded sense can lead to preferences for worse states of affairs, since it can lead to preferences for states of affairs in which the agent has more opportunity to make a positive difference to the outcome (for example, a preference that more severe disasters occur).

The latter preference might be unproblematic, for reasons related to the fact that it cannot be action-relevant. However, in the context of uncertainty, some common patterns of preference that are rationalised by a difference-making mindset run contrary to pure benevolence in ways that *are* action-relevant. This is true of both (what we called) "difference-making risk averse" and "difference-making ambiguity averse" preferences. Preferences of both these types lead to (what we called) collective defeat of goodness. Additionally, difference-making risk averse preferences violate stochastic dominance with respect to goodness, and in that sense fail to track the (ex ante) betterness relation, even in the context of individual action.

In all these cases, it is tempting to rationalise preferences of the types we have discussed via appeal to a desire for meaning in one's life. Insofar as the route to meaning was supposed to be via engaging in benevolence, however, this motivation for difference-making preferences seems in the end of dubious coherence.

It is not inevitable that a preference to make a better difference leads to any of the phenomena we have discussed here. If one prefers to make a better difference in the outcome-comparison sense, and if in addition any aversion to risk or ambiguity is only with

respect to overall good, then one will avoid all of the pitfalls that we have discussed. That is, however, the case in which one's preference to make a better difference is extensionally equivalent to a preference that the outcome be better, at least on the matter of dispositions to action. Wherever the two come apart, we have suggested, difference-making preferences run into trouble.

Nor are these points of merely academic interest. Moving away from difference-making and towards goodness-increasing preferences has the potential to significantly revise many common patterns of would-be benevolent behaviour. The common tendencies towards diversifying one's individual philanthropic portfolio, and neglecting mitigation of extinction risks in favour of "safer"-seeming near-termist options, are just some examples of this.

References

Jonathan Bennett. *The Act Itself*. Clarendon Press, Oxford, 1995.

Lara Buchak. *Risk and Rationality*. Oxford University Press, 2013.

Lara Buchak. How should risk and ambiguity affect our charitable giving? GPI working paper 8-2022, 2022.

Todd Calder. Against consequentialist theories of virtue and vice. *Utilitas* 9 (2):201–219, 2007.

Julia Driver. *Uneasy Virtue*. Cambridge University Press, Cambridge, 2001.

Brian Duncan. A theory of impact philanthropy. *Journal of Public Economics*, 88(9–10):2159–2180, 2004.

Jon Elster. The Valmont effect: The warm-glow theory of philanthropy. In Illingworth , Wenar and Pogge (2011).

Philippa Foot. The problem of abortion and the doctrine of double effect. *Oxford Review*, 5:5–15, 1967. Reprinted in Philippa Foot, *Virtues and Vices*, Oxford: Basil Blackwell, 1978.

Hilary Greaves. Evidence, cluelessness and the long term. Talk delivered at EA Student Summit 2020, 25 October 2020. Available online from https://www.youtube.com/watch?v=fySZIYi2goY . Centre for Effective Altruism, 2020.

Hilary Greaves and William MacAskill. The case for strong longtermism. GPI Working Paper No. 5–2021, 2021.

Hilary Greaves, William MacAskill, and Andreas Mogensen. Risk aversion, ambiguity aversion and longtermism. Unpublished manuscript, n.d.

Josef Hadar and William R. Russell. Rules for ordering uncertain prospects. *The American Economic Review*, 59(1):25-34, 1969.

Thomas Hurka. *Virtue, Vice, and Value*. Oxford University Press, New York, 2001.

Patricia Illingworth, Lief Wenar and Thomas Pogge (eds), 2011. *Giving Well: The Ethics of Philanthropy*. Oxford University Press.

Petra Kosonen. *Tiny probabilities of vast value.* DPhil thesis, University of Oxford, 2022.

Richard Pettigrew. Effective altruism, risk, and human extinction. GPI working paper 2-2022, 2022.

Shelly Kagan. *The Limits of Morality*. Clarendon Press, Oxford, 1991.

Frances Kamm. Responsibility and collaboration. *Philosophy and Public Affairs*, 28(3):169–204, 1999.

William MacAskill. Replaceability, career choice, and making a difference. *Ethical Theory and Moral Practice*, 17(2):269–283, 2014.

William MacAskill. *Doing Good Better: Effective Altruism and a Radical New Way to Make a Difference*. Guardian Faber, London, 2015.

Bradley Monton. How to avoid maximizing expected utility. *Philosophers' Imprint*, 19(18):1–25, 2019.

Liam B. Murphy. *Moral Demands in Nonideal Theory*. Oxford University Press, New York, 2000.

Ted O'Donoghue and Jason Somerville. Modeling risk aversion in economics. *Journal of Economic Perspectives*, 32(2):91-114, 2018.

W. D. Ross. *The Right and the Good*. Clarendon Press, Oxford, 1930. Reprinted 2002.

Thomas Schelling. *The Strategy of Conflict*. Harvard University Press, 1960.

Peter Singer. *The Most Good You Can Do: How Effective Altruism Is Changing Ideas About Living Ethically*. Yale University Press, New Haven, Connecticut, 2015.

James Snowden. Should we give to more than one charity? In Hilary Greaves and Theron Pummer, editors, *Effective Altruism: Philosophical Issues*, chapter 5, pages 69–79. Oxford University Press, Oxford, 2019.

Ralph Wedgwood. Intrinsic values and reasons for action. *Philosophical Issues* 19 (2009): 342-363.

Bernard Williams. A critique of utilitarianism. In J.J.C. Smart and Bernard Williams, editors, *Utilitarianism: For and Against*, pages 75–150. Cambridge University Press, Cambridge, 1973.

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

Fiona Woollard and Frances Howard-Snyder. Doing vs. Allowing Harm. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021. Accessed September 2021.