What should we do with regard to climate change given that our choices will not just have an impact on the well-being of future generations, but also determine who and how many people will exist in the future?

There is a very rich scientific literature on different emission pathways and the climatic changes associated with them. There are also a substantial number of analyses of the long-term macroeconomic effects of climate policy. But science cannot say which level of warming we ought to be aiming for or how much consumption we ought to be prepared to sacrifice without an appeal to values and normative principles.

The research program Climate Ethics and Future Generations aims to offer this kind of guidance by bringing together the normative analyses from philosophy, economics, political science, social psychology, and demography. The main goal is to deliver comprehensive and cutting-edge research into ethical questions in the context of climate change policy.

This volume showcases a first collection of eleven working papers by researchers within the program, who address this question from different disciplines.

Find more information at climateethics.se.

INSTITUTE FOR FUTURES STUDIES

Box 591, SE-101 31
Stockholm, Sweden

Phone:
+46 8 402 12 00

E-mail:
info@iffs.se

Vol. 2

STUDIES ON CLIMATE ETHICS AND FUTURE GENERATIONS

Editors: Paul Bowman Katharina Berndt Rasmussen

WORKING PAPER SERIES · Vol. 2 · 2020:1-11 ·

STUDIES ON

CLIMATE ETHICS

AND FUTURE GENERATIONS

Working paper series 2020:1–11

Editors: Paul Bowman & Katharina Berndt Rasmussen

Institute for Futures Studies

Studies on Climate Ethics
and Future Generations
Vol. 2

Studies on Climate Ethics and Future Generations Vol. 2

Editors: Paul Bowman
Katharina Berndt Rasmussen

Institute for Futures Studies
Working Papers 2020:1-11
Stockholm 2020

The Institute for Futures Studies is an independent research foundation financed by contributions from the Swedish Government and through external research grants. The institute conducts interdisciplinary research on future issues and acts as a forum for a public debate on the future through publications, seminars and conferences.

Cover: Matilda Svensson
Cover image: Unsplash/Jason Shuller
Distribution: The Institute for Futures Studies, 2020

Contents

Preface 5

Person-affecting and non-identity
Krister Bykvist 11

What Is the Right Way to Make a Wrong a Right?
M.A. Roberts 39

Getting Personal – The Intuition of Neutrality Re-interpreted
Wlodek Rabinowicz 59

Persson's Merely Possible Persons
Krister Bykvist & Tim Campbell 91

Liability for Emissions without Laws or Political Institutions
Göran Duus-Otterström 105

Duties of Corrective Justice and Historical Emissions
Paul Bowman 131

The distinct moral importance of acting together
Katie Steele 159

Does Climate Change Policy Depend Importantly on Population Ethics? Deflationary Responses to the Challenges of Population Ethics for Public Policy
Gustaf Arrhenius, Mark Budolfson & Dean Spears 169

Population ethics and the prospects for fertility policy as climate mitigation policy
Mark Budolfson & Dean Spears 199

Climate change denial among radical right-wing supporters
Kirsti M. Jylhä, Pontus Strimling & Jens Rydgren 219

How Much Do We Value Future Generations? Climate Change, Debt, and Attitudes towards Policies for Improving Future Lives
Malcolm Fairbrother, Gustaf Arrhenius, Krister Bykvist & Tim Campbell 237

Preface

This volume comprises the second round of preprint papers written as part of the Climate Ethics and Future Generations project. This multi-disciplinary project, led by PI Gustaf Arrhenius and co-PIs Krister Bykvist and Göran Duus-Otterström, aims to provide comprehensive and cutting-edge research on ethical questions concerning future generations in the context of climate change policy. The project began in 2018 and will run through 2023, and is generously financed by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences). For more information about the Climate Ethics and Future Generations project, please visit climateethics.se.

The eleven papers in this volume are organized according to the project's three main themes: *Foundational questions in population ethics*, which concerns how we should evaluate future scenarios in which the number of people, their welfare, and their identities may vary; *Climate justice*, which concerns the just distribution of the burdens and benefits of climate change and climate policy, both intra- and intergenerationally; and *From theory to practice*, which concerns how to apply normative theories to the circumstances of climate change, in light of both normative uncertainty and practical constraints.

The first four papers in this volume belong to the first project theme; each paper examines an important theoretical question in population ethics. The volume's first paper is by co-PI Krister Bykvist and sets out to investigate a tension between a common formulation of the person-affecting constraint (according to which "what is better (worse) must be better (worse) for someone") and our considered judgments in many non-identity cases – for instance, the judgment that creating a miserable person would make the world worse. Bykvist considers a number of recent attempts to resolve this tension, but finds all of them unsuccessful. Bykvist ultimately concludes that we should reject the common formulation of the person-affecting constraint.

The volume's second contribution, by Melinda Roberts, is also concerned with non-identity cases. According to Roberts, many of the most challenging versions of the non-identity problem inherently involve claims about probability. That fact in itself might suggest that the underlying non-identity cases should be evaluated using expected value theory. Noting that expected value theory has serious problems, Roberts describes and evaluates an alternative vehicle, the concept of *probable value*, for bringing considerations of probability to bear in our analysis of the problem cases. She concludes that the probable value approach, like the expected value approach, easily accommodates the result that the choices that we

consider wrong in the relevant non-identity cases in fact do make things worse for particular existing and future people.

Wlodek Rabinowitz is the author of the volume's third contribution, which examines what Rabinowitz calls the "Intuition of Neutrality." According to Rabinowitz's formulation, the Intuition of Neutrality holds that "there is a range of wellbeing levels such that adding people with lives at these levels doesn't make the world either better or worse." Noting that the Intuition of Neutrality appears to be in conflict with a tenet of welfarism (the tenet that what is good for a person is impersonally good), Rabinowitz explores the implications of a position he considered in an earlier work that was intended to resolve the conflict. This position involves a significant re-interpretation of the Intuition of Neutrality; it gets restricted to the wellbeing levels of lives that are personal neutral, i.e. lives that are neither good nor bad for persons.

Rounding out papers that reflect the *Foundational questions* theme is one by co-PI Krister Bykvist and Tim Campbell. Their contribution discusses a recent argument by Ingmar Person for the seemingly paradoxical claim that things can be better (worse) for a person even in a world in which the person does not exist. Persson thinks he can argue for this claim from 'incontestable' premises. Bykvist and Campbell show that this is far from true. They also argue, against Persson, that it is possible to make sense of our obligations to future generations without letting merely possible beings into the moral club.

The next two papers, by Göran Duus-Otterström and Paul Bowman respectively, fall squarely under the project's second theme, climate justice. Both Duus-Otterström and Bowman examine arguments that aim to limit the scope of the so-called "polluter pays principle" – roughly, the principle that those agents who have produced excessive emissions are morally liable to bear the burdens of addressing climate change.

In his contribution, Duus-Otterström considers the view that moral liability for emissions presupposes the existence of a just system of legal regulation of emissions. Duus-Otterström argues that the view fails to account for the fact that agents have a moral duty to promote the emergence of a just system of legal regulation of emissions. Because the production of excessive emissions makes the emergence of such a system less likely, contrary to what some critics have argued, agents can be morally liable for their pre-legal emissions.

For his part, Bowman considers a different argument for the claim that actors are morally liable only for their recent emissions. According to this argument, agents were, until relatively recently, non-culpably ignorant of the fact that their emissions were causing harmful climate change. While Bowman accepts that non-culpable ignorance of the harmful effects of one's action normally defeats moral

liability to bear the costs of rectifying these harmful effects, Bowman argues that there is an important exception to this general principle: namely, when the agent would have performed a relevantly similar action had the agent not been non-culpably ignorant of these effects. Bowman then argues that it is plausible that many agents who produced excessive emissions would have done so even had they known these emissions would contribute to harmful climate change; hence, it is plausible that they can be morally liable for these emissions.

It is highly plausible that agents have a moral reason to cooperate to bring about morally good outcomes, like those that achieve climate justice. In her contribution, Katie Steele engages with a recent argument by Garrett Cullity that even if we each barely make a difference to efforts to mitigate climate change, we nonetheless have a fundamental moral reason to join or cooperate in these efforts. Drawing on the game theoretic notion of 'team reasoning', Steele provides an account of the reason to cooperate that supplements Cullity's own account.

The final four papers in the volume reflect the project's third theme, *From theory to practice*. The contribution from PI Gustaf Arrhenius and co-authors Mark Budolfson and Dean Spears considers whether seemingly intractable problems in population axiology (the theory of the value of populations) means that we must be ignorant about which climate policies to pursue, given that different climate policies will result in populations in which the number of people, their welfare, and their identities may vary. Arrhenius, Budolfson, and Spears suggest a deflationary response to this worry in which they argue that in spite of the problems in population axiology, scepticism about climate policy in light of these problems may be unwarranted.

Budolfson and Spears co-author the next contribution in the volume. In their article, they consider whether policies that limit fertility can be an effective strategy for mitigating climate change. They argue that, contrary to what some policy debates assume, even very ambitious fertility policies would only have a very modest effect on population size in the coming decades, because of a demographic process called "population momentum." As a result, fertility policy is unlikely to have a large effect on greenhouse gas emissions, even over a span of several decades. They conclude, therefore, that, because climate policy is urgent, fertility policy is unlikely to be an effective means of mitigating climate change.

The volume's final two contributions use survey data to examine people's beliefs and attitudes on climate change and the value of future generations. In their paper, Kirsti M. Jylhä, Pontus Strimling, and Jens Rydgren investigate the relationship between climate change denial and political party affiliation, focusing especially on how the radical-right Sweden Democrats compare with two other parties in Sweden. Their analysis suggests that certain psychological factors more prevalent among

those who affiliate with mainstream and/or radical right-wing parties are a better predictor of climate change denial than party affiliation as such.

The volume's final piece, by Malcolm Fairbrother, PI Gustaf Arrhenius, co-PI Krister Bykvist, and Tim Campbell examine people's attitudes towards future generations, particularly as these attitudes bear on their support for climate change policies. Using survey data from individuals in four countries, the authors find that while most people claim to be willing to bear costs to benefit future generations, they generally claim not to be willing to support government policies that aim to tackle climate change or the national debt. According to the authors, people's aversion to future-oriented policies may not be due as much to their discounting the value of future generations per se, as it is due to their distrust of government and its ability to deliver the policies' putative benefits.

We are pleased to be able to share these fascinating and timely papers with you. We look forward to seeing what new research will emerge from the Climate Ethics and Future Generations project in the years to come.

Paul Bowman & Katharina Berndt Rasmussen
Editors

Krister Bykvist[1]

Person-affecting and non-identity[2]

According to a popular version of the person-affecting idea of morality, what is better (worse) must be better (worse) for someone. However, there seems to be a clear tension between this idea and some of our considered judgements about cases in which the existence of future people is contingent on our choice. For example, we want to say that creating a very unhappy person makes the world worse, other things being equal. In order to comply with a person-affecting morality in this case, we need to show that coming into existence can be worse for a person, but it does not seem plausible to say that it can be worse for a person to exist than not to exist. This paper discusses some recent attempts to ease this tension, and it is argued that none of these attempts is convincing. That leaves us with only one option: to reject the person-affecting constraint in its current form.

[1] Institute for Futures Studies & Department of Philosophy, Stockholm University, krister.bykvist@iffs.se.
[2] Financial support by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences) is gratefully acknowledged.

1. Introduction

One important part of morality is concerned with what is better or worse for people. According to a popular version of this *person-affecting* idea of morality, what is better (worse) must be *better (worse) for* someone.[3] However, it is unclear how we are to put this idea to use in non-identity cases, i.e., cases where, depending on what we decide to do, different people will come to exist in the future. Indeed, there seems to be a clear tension between the person-affecting idea and some of our considered judgements about non-identity cases. In at least some non-identity cases we want to say that one outcome is better (or worse) than another in virtue of the wellbeing of people who do not exist in both. For example, we want to say that creating a very unhappy person makes the world worse, other things being equal. But how can we say this, if an outcome is worse only if it worse for someone? In order to comply with a person-affecting morality in this case, we need to show that coming into existence can be worse for a person. But can it really be worse for a person to exist than not to exist, and thus better for her not to exist than to exist? That seems to require that the person would have been better off not existing, which sounds paradoxical.

In this paper, I am going to discuss some recent attempts to ease this tension. According to these attempts, we can stick to a person-affecting morality and still avoid the counterintuitive judgement that no outcome is better or worse in virtue of the wellbeing of people whose existence is contingent on our choice. I shall show that none of these attempts is convincing. That leaves us with only one option: to reject the person-affecting constraint in its current form.

In section 2, I shall say more about non-identity cases, and list the most morally salient ones. In section 3, I shall make more precise what a person-affecting morality amounts to. In section 4, I shall present an argument that spells out the tension between person-affecting morality and our judgements about non-identity cases. The argument's conclusion is that no outcome can be better or worse than another in terms of the well-being of people who do not exist in both. In sections 5 to 10, I shall discuss possible ways to resist this argument while sticking to a person-affecting morality. I shall especially focus on the approach recently defended by the so-called 'Scandinavian existentialists'.[4] I shall argue that the main problem with their approach is that they fail to fully acknowledge what it means to say that an abstract state of affairs has value.

[3] See Temkin (1993a), (1993b), and Holtug (1996). The label "Person-Affecting Restriction" was introduced by Glover (1977), p. 66, but see also Narveson (1967).
[4] See, for instance, Arrhenius & Rabinowicz (2010), (2015), Johansson (2010), and Holtug (2001). See also, Adler (2009), and Adler (2011) for similar ideas. Some seeds for this approach seem to have been planted already in Parfit (1995), appendix G, p. 490.

2. Non-identity cases

Non-identity cases are cases where the alternative outcomes do not contain the same people, but they might contain the same number of people. There are many different kinds of non-identity cases, but there are five groups that are especially morally salient (below 'good/bad life means 'good/bad for the person leading the life'):

> *Good lives versus no lives*: The mere addition of a number of people with good lives, or no addition at all.
>
> *Bad lives versus no lives*: The mere addition of a number of people with bad lives, or no addition at all.
>
> *Good lives versus bad lives*: The mere addition of a number of people with good lives or the mere addition of a number of different people with bad lives.
>
> *Good lives versus even better lives*. The mere addition of a number of people whose lives are good for them or the mere addition of a number of different people with better lives.
>
> *Bad lives versus even worse lives*. The mere addition of a number of people with bad lives or the mere addition of a number of different people with worse lives.

There is no general agreement on how to assess these different mere additions. For example, some think that in the case of *Good lives versus no life*, a mere addition of a good life always makes the world better, whereas others deny it. But there is a wide agreement that not every case is a matter of indifference (or incomparability); some of these additions do make the world better or worse. For example, most would agree that in case of *Good lives versus bad lives*, a mere addition of huge number of very good lives (all equally good) makes the world better than a mere addition of a huge number of very bad lives. Similarly, in the case of *Bad lives versus no lives*, most would agree that the mere addition of a huge number of very bad lives is worse than no addition at all.

3. The person-affecting constraint

A common formulation of the person-affecting constraint is this:

> *The person-affecting constraint (PAC)*:
>
> If A is better (worse) than B, then A is better (worse) than B for someone.

This formulation calls out for some clarifications. First, it is assumed that PAC is necessarily true and that A and B are *abstract alternatives*, in the sense that they can exist without being instantiated, but they cannot both be instantiated. It is important that A and B are alternatives that cannot both be instantiated, since we are comparing alternatives in which some people exist with alternatives in which these people do not exist. I shall accept these assumptions but not take a stand on the exact ontological nature of the alternatives and the instantiation relation. The alternatives can be seen as abstract *states of affairs* that can exist without *obtaining or being realized*, *properties* that can exist without being *exemplified*, *event-types* that can exist without being *tokened*, or *propositions* that can exist without being *true*. For ease of exposition, I am going to use 'exemplify' as a catch-all term for these more specific instantiation relations.

Second, to make sure we have all the information needed to make overall assessments of the alternatives, I shall assume that they are *consistent* and *complete* with respect to what matters for well-being (including enablers and disablers, if there are such things). At times, I shall use 'S's existence' ('S's non-existence') as a convenient way of referring to any abstract complete and consistent alternative that includes (precludes) the existence of S, in the sense that, necessarily, if it were exemplified, S would (not) exist.

Third, I am bracketing issues about non-welfarist values, since I focus on the part of morality that is concerned with well-being, 'benevolence', as we might call it. But I also bracket the value of inequality of wellbeing and other welfarist 'pattern' values. Without this bracketing PAC would be controversial even in cases where the identity of people is not at stake. For example, if one gives a lot of weight to inequality of wellbeing, making someone worse off need not make things worse overall, since this can be a way to reduce inequality (by 'levelling down') without too much of a sacrifice in total wellbeing.

Fourth, there is an obvious problem with this formulation of PAC, if the quantifier 'for someone' ranges only over *actual* people. Suppose A and B contain only *non-actual* individuals, i.e., individuals that do not exist in the actual world, and that in A

everyone is extremely happy (and equally so) and in B everyone is extremely unhappy. Then PAC entails that A is not better than B, for there is no actual person for whom A is better than B.

A formulation that avoids this problem is this one:

> If A is better (worse) than B, then A is better (worse) than B for someone *who exists in A or in B*.[5]

More precisely, unpacking the disjunctive 'in'-locution:

> If A is better (worse) than B, then either A would better (worse) than B for someone, if A were exemplified, or A would be better (worse) than B for someone, if B were exemplified.

Finally, I shall assume that A is better for S than B if and only A, S, and B exist and are standing in the relation expressed by '__is better for__than__'. Clear evidence for this is that 'A is better for S than B' entails (a) there is *something* that is better for S than B, (b) A is better for S than *something*, and (c) A is better than B for *someone*.

4. The tension between PAC and our intuitive judgements about non-identity cases

Here is an instructive argument for the claim that if the person-affecting constraint is true, then there is no non-identity case of the mere addition kind listed above in which one alternative is better or worse than another and in which. Let A and B be two alternatives in such a case, so

> (i) A and B differ with respect to the identity of people, but are equally good for those who exist in both.[6]

Now we add PAC

> (ii) If X is better (worse) than Y, then X is better (worse) than Y for someone who exists in X or in Y.

[5] The need for this reformulation is pointed out in Holtug (2001).
[6] Alternatively, we can assume, a bit unrealistically, that A and B do not share any individuals.

And two plausible general principles:

> (iii) If X is better (worse) than Y for a person who exists in X or in Y, then she would have been better (worse) off in X than in Y (*Better-for entails better-off*).[7]

> (iv) No one can be better off (worse off) existing than not existing. (*Well-being entails being*).

which means that

> (v) No one in A or in B is better (worse) off in A than in B. (from (i) and (iv))

> (vi) A is not better (worse) than B for anyone in A or in B. (from (iii) and (v))

And we get the conclusion

> A is not better (worse) than B. (from (ii) and (vi))

Since we chose A and B arbitrarily, this conclusion shows that there is *no* non-identity case of the mere addition kind in which one alternative can be said to be better than another. But this is very worrying, since, as I pointed out in section 2, we would like to say that, at least in some non-identity cases, one alternative is definitely better than another. For example, the mere addition of a huge number of very good lives (all equally good) is definitely better than the mere addition of a huge number of very bad lives.

Recently, there have been attempts to resist this conclusion while holding on to PAC. Some (Roberts, Voorhoeve, and Fleurbaey) reject *Well-being entails being* and claim that non-existence does not preclude being better off (or worse-off).[8] Others (Adler, Arrhenius, Rabinowicz, Holtug, Johansson) instead reject *Better-for entails better-off* and claim that existence can be better for a person than non-existence even though the person would not be better off existing than not existing. A third option is to deny (i), i.e., deny that there are any non-identity cases, because one thinks that in all the worlds in which a person is not conceived, she still exists as a

[7] Adler (2009), p. 1503, also states this principle, but he rejects it.
[8] Roberts (2015) does not defend PAC, but a weaker principle she calls 'the person-based intuition', according to which an outcome A is worse than an outcome B only if A is worse for someone than some alternative outcome Z, where Z need not be identical to B. However, she would have to defend PAC, if it is restricted to cases where A and B are the only available outcomes.

merely possible person, who has wellbeing. I shall argue that none of these ways of blocking the argument works. This leaves PAC itself as the only remaining culprit.

5. Against *Well-being entails being*: Voorhoeve and Fleurbaey

Voorhoeve and Fleurbaey claim that there are non-existent persons.[9] This claim smacks of incoherence, since it is not easy to find a coherent interpretation of 'There are persons who do not exist'. After all, to be a person, at least on the most natural interpretation of 'person', requires existence. They could drop the reference to persons and just say there are *individuals* who do not exist. But this is not much better, since 'there are' seems to have existential import: to say that there are individuals who do not exist is to say that there *exist* individuals who do not exist.

Fleurbaey and Voorhoeve agree with *part* of spirit of the principle *Wellbeing entails being,* for they accept that non-existent individuals have 'no wellbeing level'. But they nevertheless claim that non-existent individuals can be better-off or worse-off. More exactly, they claim that if S exists in B but not A, it can still be true that if A *were* the case, then A *would* be better for S than B. But note that S is better off in A than in B, if it is both true that A would be better for S than B if A were the case, and B would be worse for S than A if B were the case.[10] How can one accept that one can be better off never existing without assigning well-being levels to the never existing? Their answer:

> (...) one can sensibly hold that a particular life can be better for a person than never existing without assigning a level of well-being to never existing. It is sufficient that there is a level of wellbeing, when existing, that is deemed *equivalent* to never existing. (Call this the 'personal-value indifference' value of wellbeing: it is often referred to as the 'neutral level'). Then, we submit, enjoying a greater wellbeing than this level implies that a person's life is better for her than not existing.[11]

Even though Voorhoeve and Fleurbaey state that the conclusion is that a person's life is better for her than not existing, it is clear from the context that they think it also establishes the claim that a person would be *better off* existing and leading the live than not existing.

[9] Voorhoeve and Fleurbaey (2015), p. 98–100.
[10] Voorhoeve and Fleurbaey (2015), p. 100.
[11] Ibid.

What does 'deemed equivalent' mean? Deemed equivalent in terms of what? They claim that if a life L has a higher level of well-being than a life L* and a life L* is 'deemed equivalent' as non-existence, then L is better for the person than non-existence. Surely, for this to work 'deemed equivalent' must mean 'truthfully deemed equivalent *in value* for the person', or, more succinctly, 'equally as good for the person as'. But then if L* has a certain level of wellbeing for a person and L* is equally as good for her as her non-existence, then her non-existence also has the same level of well-being. In general, if A has a certain level of well-being for S, and A is equally as good for S as B, then B has the same level of well-being for S as A. Indeed, this holds for other comparatives too: if I have a certain height and you are as tall as me, then you also have the same height. If I have a certain weight and you are as heavy as me, then you have the same weight. So, I can't see how Voorhoeve and Fleurbaey can avoid assigning a level of wellbeing to non-existence.

In any case, their account directly runs into problems with the following principle, which I take to give a defining feature of better off:

> *Wouldy Better-off*
>
> S is better off in A than in B iff the value A would have for S, if A were exemplified, is greater than the value B would have for S, if B were exemplified.[12]

This principle is very plausible, since, as Broome has pointed out, to be better off in A than in B is not just to stand in a relation to A and B; it involves standing in relations *in* A and *in* B.[13] But, I would like to add, to have stand in a relation *in* X, where X is a complete and consistent abstract alternative, must be understood *counterfactually*: one would stand in the relation, if X were exemplified. Since nothing can have a value for a non-existent person – a claim Voorhoeve and Fleurbaey themselves come very close to endorsing, since they say that non-existent persons have no level of wellbeing - the above principle rules out their account.[14]

---

[12] As I will explain in section 8, it is not just that A would have value for S, if A were exemplified; it is also true that some part of the world would also have value for S in virtue of exemplifying A.

[13] Broome (2004).

[14] It is true that it is popular to say things like 'I would have been better off never existing', and 'I would have been better off if I had never been born', but this must be understood as *hyperbolic idioms*. Taken literally, they are clearly false, if existence is understood in its minimal generic sense of being something, i.e., being identical to something. After all, it is not popular to say 'I would have been better off not being anything, i.e., not being identical to anything'. In section 7, I shall show that they are false even if we accept that I could have existed as a merely possible person, if I had not been born, or brought into a concrete existence. In any case, we know what is *communicated* by these false idioms: I am really *badly off* existing, which is a claim that does not require any wellbeing comparisons between existing and not existing.

6. Against *Well-being entails being*: Roberts

Roberts accepts that a person can only have properties in situations in which she exists, but she maintains we can still say that a person can have zero level of well-being in situations in which she does not exist. She maintains that:

> (...) Nora does not have any properties at all at any alternative at which she does not exists and (...), where Nora has no properties at all, all the properties that she does have – the empty set – add up to zero level of wellbeing.

And this means, she claims, that we can say that the person better off leading a good life than not existing at all. Now, it is not clear what she means by the empty set of properties 'adding up' to zero level of wellbeing. Obviously, she cannot mean that to have zero level of wellbeing is the *same* as not having any properties, for I can have zero level of wellbeing and still have properties, for instance, by being unconscious and not having any good or bad things happening to me, or, alternatively, by having good and bad things happening to me when the good things exactly balances the bad things.

A recent argument for her claim that one can be better off existing than not existing is her 'zero money - zero wellbeing' analogy. Roberts suggests that to say that a person has zero wellbeing is analogous to saying that she has zero money. Since you can have zero money in a country in which you do not live (China, for instance, in Robert's case), you can also have zero wellbeing in a world in which you do not exist. The analogy is not convincing.[15] To say that one has zero money in China is to say that one has *no* money in China. By analogy to say that one has zero wellbeing in a world in which one does not exist is to say that one has *no* wellbeing in this world, neither positive, negative, nor neutral wellbeing. But then we cannot say that the person is better off existing than not existing. For, as pointed out in the previous section, 'better off' is *wouldy* in the sense that to be better off in A than in B requires that one A would have a certain value for one, if A were realized, and also that B would have a certain value for one, if B were realized.

But perhaps Roberts assumes that to say that S has neutral wellbeing is just to say that it is *not the case* S has overall positive or negative wellbeing. This is one way to understand her claim that Nora's empty set of properties somehow 'adds up' to a zero level of wellbeing. But this is not true. Lots of things lack positive or negative wellbeing, but it is not true to say that they therefore must have a neutral level of well-being. Abstract entities such as number 2, for instance, or certain concrete things, such as my socks, do not have any wellbeing whatsoever.

[15] For a different criticism of Roberts' analogy, see Johansson (2010), p. 289.

Note that this is not a quibble about terminology. To have neutral wellbeing is evaluatively and normatively significant in a way that lacking any wellbeing whatsoever is not. If a person has neutral wellbeing, we can talk about the *equality* and *inequality* between her wellbeing and others', and we can say that she is *better off* and *worse off* than others. If she has neutral well-being, then if we make her better off, things will be *good* for her, and if we make her worse off, things will be *bad* for her. Furthermore, the fact that she has neutral wellbeing has normative implycations. For example, it is then fitting to take a neutral attitude towards her situation. It is also fitting to feel *sympathy* for her, when, through no fault of her, she ends up having only neutral wellbeing in a situation in which she could have had great positive well-being. None of these implications hold when a person lacks wellbeing altogether.

It should be noted that in deciding whether it makes sense to say that one is better off existing than not existing it is easy to be misled by locutions such as

> There is more value for S in A than in B.
>
> There is more well-being for S in A than in B.

for these sentences can be true because (respectively):

> There is some value for S in A but *none* in B, for S does not exist in B.
>
> There is some wellbeing for S in A and *none* in B, for S does not exist in B.

The same phenomenon shows up in other contexts. There is more water in the ocean than in the dry desert, since there is some water in the ocean and none in the dry desert. There is more pain in this world than in a world in which there are no sentient beings, since there is some pain in our world and none in sentience-less world. Or to take an example that is closer to Roberts' own example: I have more money in Sweden than in China, for I have some money in Sweden and none in China. In general, there is a sense of 'more F' that can be used to state that there is more F in X than in Y, when there is some F in X and none in Y.

But 'more wellbeing' and 'more value for', in this sense, should not be conflated with 'being better off'. If A is good for S and S has no wellbeing in B, because S does not exist in B, then, S cannot be better off in A than in B, for this would imply the false claim that B would have a value for S, if B were exemplified.

Roberts' insistence that non-existent persons can have zero level of wellbeing

might thus be based on a conflation between two different readings of 'zero level of wellbeing', namely, 'no wellbeing' and 'neutral wellbeing'.[16] Roberts is right to say that there is more wellbeing for me in an alternative A, in which I exist and have a good life, than in an alternative B, in which I do not exist. But this does not show that B would be neutral for me, if B were realized, which is the claim she needs to establish in order to show that I would be better off in A than in B.

7. Against genuine non-identity cases: necessarily existing merely possible persons

A more radical reaction to the argument is to deny that there are any genuine non-identity cases, where the outcomes differ in terms of who exists in them.[17] If you had not been conceived, you would still have existed but only as *a merely possible person*.[18] Furthermore, the daughter you could have had but in fact did not have, still exists as a merely possible person. More generally, if a person exists in a world, as a person, then she exists, as a merely possible person, in all worlds in which she is not conceived. So, there cannot be any non-identity cases, for every person exists necessarily, (but not necessarily as a person).

A merely possible person is something that is in fact not a person but could have been a person.[19] In general, if x is a merely possible F, it does not follow that x is F. Note that a merely possible person is not the same as a *potential* person. A potential person is a *concrete* thing – for example, a fertilized egg - that is not a person but can *develop* into a person. A merely possible person is a non-concrete thing that could have been a person (and concrete). I am not identical to the fertilized egg that developed into me, but I am identical to some merely possible person in a world in which I am not conceived. Of course, that there are necessarily existing merely possible persons in this sense is a very controversial metaphysical assumption, but let this pass and assume for the sake of the argument that there are such beings.

The important question is whether merely possible persons can have wellbeing.

---

[16] Holtug (2001) seems to make a similar conflation: he equates having nothing good or bad happening to you with having neutral wellbeing.

[17] Williamson (2013), p. 63, suggests that this view could have important ramifications for ethics: if one believes in necessarily existing merely possible persons one cannot simply dismiss as meaningless the claim that it would have been better for a person not to have been born.

[18] This is assumed without arguments in Hare (1988), p. 281.

[19] I am here ignoring David Lewis' alternative reading of 'merely possible persons'. On his infamous realist view of possible worlds, presented in Lewis (1986), all possible worlds are concrete, but they are causally and spatiotemporally isolated from each other. To be a merely possible person in a possible world w is to be a concrete person that does not exist in w, but who exists in a *different* possible world. This alternative reading of 'merely possible person' will not be of any help in easing the tension between PAC and our judgements about non-identity cases, for someone who is a merely possible person in a possible world w, in this Lewisian sense, does not exist in w and thus cannot have any wellbeing in w.

If they cannot have wellbeing, then we cannot evade the tension between PAC and our considered judgements about some that some purported non-identity cases, for example, that it is better to create a very happy person than to create a very unhappy person. It does not help to be told that the happy person would still have existed as a merely possible person, had she not been conceived, if she then would have lacked any wellbeing whatsoever. If she would have lacked any wellbeing whatsoever, then we can still not say that she would have been better off conceived than not conceived. For, as *Wouldy Better-off* demands, if you are better off conceived than not conceived, then you would have had some wellbeing, if you had not been conceived.

So, the crucial question is 'Is it possible for a merely possible person to have well-being?' Or, more exactly, 'Is it possible that: there is a merely possible person, who has well-being?' We are not interested in the *narrow scope* reading of the question: 'Is there is a merely possible person such that it is possible that she has well-being?', for an affirmative answer to this question does not help. In order to truthfully say that having been unborn *would* have been worse for you than having been born we would need to assign wellbeing to you as an unborn merely possible person. It is not enough to say that you, as a merely possible person, could have had a well-being, because you would have had it, *if* you had been born.

It is very doubtful that you could have had wellbeing, if you had not been conceived. For if you had not been conceived, you would have been a merely possible person, a merely possible animal, a merely possible human, a merely possible philosopher, ... Why should we not continue this list with 'a merely possible bearer of *well-being*? After all, it is very plausible to think that having well-being requires being concrete in some way, for example, having a mind, body, being in space or time, or having causal power.

One could perhaps object that to have well-being it is enough that one has the *capacity* to be concrete in a certain way, having a mind, say.[20] Since merely possible people, if they exist, are such that they possibly have minds (they have minds in all worlds in which they are born and develop into mature persons), they also have the capacity to have a mind. So, if you had not been born, you would have been a merely possible person with the capacity to have a mind. And since a capacity to have a mind is sufficient for having well-being, you can be assigned well-being as a merely possible person.

This argument goes wrong in assuming that to have a capacity to have F it is enough that one is such that one possibly has F. This is not true. For example, I am such that I possibly jump to the moon, but I do not have the capacity to do it.[21] To have a capacity to do or have something requires more than just having the purely

[20] This argument is discussed in Bradley (2009), Chapter 3, Section 3.5.
[21] Johansson (2010) makes this important observation.

modal feature of being such that one possibly does or have it. Arguably, what is required is that the capacity is somehow grounded in features that are not purely modal.

Another argument for the possibility of assigning well-being to merely possible persons would go like this. It is true that to have neutral well-being is not just to lack the properties of having overall positive or negative well-being. To have neutral well-being is a property in its own right. But as a merely possible person you do have the *negative* property of being such that you do not have overall positive or negative well-being, and this property is just the property of having neutral well-being.

This argument overgeneralizes, however. Everything that could not have positive or negative well-being has the property of being such that it has no positive or negative well-being. For instance, the number 2, the null set, and the grain of sand in my pocket all have this property. If the argument were right, we would have to say that these things too have neutral well-being, which of course would be absurd.

One could avoid this objection by revising the definition of neutral wellbeing in the following way: to have neutral well-being is to be a *welfare bearer* (welfare subject) who has the negative property of lacking overall positive and negative wellbeing. Obviously, the number two, the null set, and the grain of the sand in my pocket are not welfare bearers, since it is inconceivable that they have any wellbeing whatsoever.

This maneuver would block the absurd conclusion, but it is important to note that it cannot be used more constructively to show that merely possible people have wellbeing. To ask whether something is a welfare bearer is just to ask whether it fulfils the requirements for having well-being and I have previously shown that it is very doubtful that they do that. Recall that they are not concrete and thus cannot have any of the properties or capacities that requires concreteness.

The prospects look bleak for finding a convincing argument for the conclusion that merely possible people could have well-being levels. They seem too 'thin', metaphysically speaking, since they only exemplify trivial or logical properties, such as being self-identical, and purely modal properties not grounded in other properties, such as being possibly F, for any F that the individual exemplifies in a world in which it is concrete.

Apart from these more metaphysical considerations, one could point out the *normatively* absurd consequences of accepting that merely possible persons have wellbeing. This objection has to do with the fact that the notion of well-being is closely tied to reasons to care and feel sympathy. When someone is made worse-off (for no faults of her own) we have reason to care and feel sympathy for her. Consider a possible scenario in which my parents decide not to have a second child and, as a consequence, I am not born. According to the account in question, I am made worse-

off by their decision. In fact, assuming that I would have had a great life, if conceived, I am made *significantly* worse-off. Now, being made significantly worse-off gives people reason to feel sympathy. So, a neutrality account implies that people in this scenario have reason to feel sympathy for me as an unborn merely possible person. But this seems absurd, so the account must be mistaken.

One reply here is to say that we only have reason to feel sympathy for people who are made worse off by being caused to *suffer*. And when my parents decided not to conceive me, they did not cause me to suffer. But this is an all too narrow understanding of sympathy. We can have reason to feel sympathy towards people who do not suffer. For example, we have reason to feel sympathy towards people who are being made unconscious and miss out on good experiences and activities.

Another reply is to say that we have reason to feel sympathy only towards people we *know* about. In the scenario in which I am not born I exist, but since no one knows about me in this scenario no one has reason to feel sympathy for me. But this reply ignores the possibility that we can know about someone just by being able to *refer* to the person. For example, consider a scenario in which my parents decide not to have intercourse, but if they had decided, I would have been conceived. Then they could know about me under the description 'the person who would have been conceived if had had intercourse now'. Of course, this is not a particularly vivid representation of me, but reasons to feel sympathy do not require vivid representations of victims (even though they make it easier to feel sympathy). Just knowing that your closest neighbor was made significantly worse off gives you some reason to feel sympathy even though you know him only under the description 'your closest neighbor'.

All in all, then, there are both strong metaphysical and normative arguments against the possibility of there being merely possible persons with well-being.

8. Against *Better-for entails better-off*: Arrhenius and Rabinowicz

Common to all Scandinavian existentialists is that they think that 'better-for' expresses a relation between abstract states of affairs and individuals. They also accept that if S exists and state of affairs p is good for S and S does not exist in state of affairs q, then S cannot be better off in p than in q; but they insist that in this case p is better for S than q.[22] So, they all deny *Better-for entails Better-off*. Since Arrhenius & Rabinowicz have developed their version of Scandinavian existen-

[22] Adler (2009), p. 1505, and Adler (2011), p. 220, makes a similar point as a defence of a preference-based account of wellbeing. He states that even though it does not make sense to say that you could have been worse off not existing, it does makes sense to say that your non-existence is worse for you than your existence in virtue of the fact that you prefer your existence to your non-existence.

tialism further than others, I shall focus on them in the following, but my criticisms apply to the whole camp.

Arrhenius & Rabinowicz point out that if S exists, then there are no missing relata.[23] We have the two abstract states of affairs p and q, and also the individual S. They also point out that it is not true in general that if a relation R holds over abstract states of affairs p, q, and person S, then R would hold over p, q, and S even if p were to obtain. For example, it does not hold if the relation is preference and the states of affairs are your non-existence and your existence and the person is you. So, the sheer logic of relations cannot be used to establish that if p is better for S than q, then p would have value for S (comparative or non-comparative), if p were to obtain.

I agree that the sheer logic of relations is not helpful here. But I think there is a strong argument against their view. Before I present the argument, we need to consider what it means, more generally, to say that an exemplifiable abstract entity has value. Let us first consider abstract *properties*. What do we mean when we say that properties such as courage and benevolence have value? At least the following, I maintain: If courage is good, then it is possible for people to be good in virtue of exemplifying courage. If benevolence is good, then it is possible for people to be good in virtue of exemplifying benevolence. The same holds for all other virtue properties. The slogan is: a good character trait could 'rate a person a plus'.

Similar things can be said about functional goodness. If sharpness is a good property for knives, then it is possible for knives to be good in virtue of exemplifying sharpness. In short, sharpness could 'rate a knife a plus'. If empathy is a good property for a counsellor, then it is possible for counsellors to be good in virtue of exemplifying empathy. In slogan form, empathy could 'rate a counsellor a plus.'

As Sven Danielsson, drawing on Chisholm, has suggested, this idea can be generalized to the value of *states of affairs*.[24] If a state of affairs p is good, then it is possible for something ('the universe', Chisholm labels it) to be good in virtue of exemplifying p. For example, if the state of affairs my feeling pleasure is intrinsically good, then it is possible that something is intrinsically good, at least to some extent, by exemplifying this state of affairs. As Chisholm puts it, the slogan is: A good state of affairs could 'rate the universe a plus'.[25]

Finally, applied to well-being the idea is that if a state of affairs p is good for you, then it is possible for something to be good for you in virtue of exemplifying p. For

[23] Holtug (2001), pp. 139-140, makes the same point.
[24] Chisholm (1966), Danielsson, unpublished manuscript.
[25] In fact, I think this idea can be generalized even further to the normative status of actions. If an abstract act-type, such as lying or killing, is wrong, then it is possible that some act-token of these types is *pro tanto* wrong in virtue of exemplifying the type in question. Indeed, I think the idea can be generalized to some *non-moral* cases too. For example, it seems natural to think that if eating vegetables is healthy, then it is possible that someone is made healthy (to some extent) by exemplifying the act-type eating vegetables.

example, if the state of affairs of my feeling pleasure is good for me, then it is possible that something is good, at least to some extent, for me in virtue of exemplifying this state of affairs. The slogan is: states of affairs that are good for you could rate something a plus for you. They are possible good-for-you makers.

It may not be perfectly obvious why this idea is a natural generalization of the idea that good properties are possible good-makers. But this generalization is difficult to resist once it is acknowledged that abstract states of affairs are similar to ordinary properties. States of affair are most plausibly seen as *ways* things could be. After all, they are states *of* something, namely, 'affairs'. If states of affairs are ways things could be, then possible worlds are maximal consistent states of affairs, i.e., *maximal* ways things could be. There is only one world, 'the totality of what exists', 'you and all your surroundings', and what we call possible worlds are more adequately called possible world-*states*, i.e., maximal ways the world could be.[26] A state of affairs p is exemplified just in case the world exemplifies a world-*state* that entails p. A state of affairs p obtains *in* a possible world-state w just in case w entails p.[27]

So far, we have considered *goodness* of abstracta that can be exemplified. But the same idea applies to badness, neutrality, or any other kind of value of such abstracta[28] The general principle, *Possible Value-making*, is that, for all exemplifiable abstracta X, if X has a certain value, then it is possible that something has this value in virtue of exemplifying X. For example, if X is good, then it is possible that something is good in virtue of exemplifying X; if it is bad, then it is possible that something is bad in virtue of exemplifying X, and if it is neutral, then it is possible that something is neutral in virtue of exemplifying X. In slogan form: an X that has value could rate some X-exemplifier this value. So, a good X could rate some X-exemplifier a plus, a bad X could rate some X-exemplifier a minus, and a neutral X could rate some X-exemplifier a zero.

The exemplifier of an abstract X is (some part) of *the* world, in the sense of 'the totality of things', or 'you and all your surroundings'. What the exemplifier consists in depends on the nature of X. The exemplifier is a token event, if X is an event-type;

[26] I take no stand on what *the world*, 'you and all your surroundings', consists of. In particular, I do not assume that it only consists of concrete individuals.

[27] States of affairs behave like properties even if one follows David Lewis and takes the notion of a concrete possible world as primitive and identify abstract states of affairs with sets of concrete possible worlds, for then states of affairs can be seen as extensional properties of concrete possible worlds. A state of affairs is exemplified just in case the actual concrete world is a member of p (i.e., the set of all the p-worlds). A state of affairs p is exemplified by a concrete possible world w just in case w is a member of p.

[28] Examples of other kinds of the value would be the indeterminate values of disjunctions that consist of disjuncts with different absolute values (e.g., being happy or being unhappy, being happy or being indifferent), and what Gustafsson (2016), p. 855, calls the 'blank' values - whatever has such a value can be worse than what is good and better than what is bad but nevertheless fail to be equally as good as what is neutral. Carlson (2016) defends a similar idea.

a particular, if it is a property; a truth-maker, if it is a proposition; and a realization of a state of affairs, if it is a state of affairs, where the realization in turn consists of 'states' i.e., properties and relations, and parts of the world (events, processes, particulars, or individuals) that exemplify these states.

This means that I am *not* talking about value-making in terms of *entailment*: one *abstract* alternative X making another *abstract* alternative Y have a certain value by being part of Y, or being entailed by X. Here are some examples of this kind of value-making:

> The property of being generous makes the complex property being generous and brave better by being part of it.
>
> The property of being sharp makes the complex property of being sharp and easy to clean better by being part of it.
>
> The state of affairs that John feels pleasure makes the complex state of affairs that John feels pleasure and Jane feels pleasure better by being entailed by it.

This kind of value-making holds between abstract alternatives and must be distinguished from the value-making I have in mind: value-making by way of *exemplification*, which holds between abstract alternatives and their exemplifiers. Here are some examples: being generous makes a *person* better, being sharp makes *a knife* better, and that John feels pleasure makes *the world* better (by making the concrete situation, in which John is feeling happiness, better). There are thus two important analogies between valuable properties and valuable states of affairs. Both are possible value-makers by way of entailment as well as possible value-makers by way of exemplification. But it is only the latter kind of value-making that is invoked in the principle *Possible Value-Making*.

To avoid misunderstandings, a few further clarifications are in place.

(i) The explanation expressed by 'something has value in virtue of exemplifying X' is *inclusive* in the sense that enablers and disablers (if there are such things) are part of the explanation. There are less inclusive notions of 'in virtue of' which would leave out the enablers and disablers and only list the factors that are enabled and not disabled to make things good, bad, or neutral.

(ii) As it stands, the principle *Possible Value-Making* is best suited to express claims about *final or intrinsic value*, since if X has this kind of value it seems clear

that it is in virtue of exemplifying X that things can have this value. [29]

(iii) The account does not say that only *concrete* things can be good in virtue of exemplifying abstract features. What exemplifies a good abstract feature can itself be abstract. So, for instance, a musical composition, assuming it to be abstract, can exemplify elegance, simplicity, and balance and be good (beautiful) in virtue of exemplifying these good abstract features.

(iv) Perhaps it is true that to say that properties and states of affairs are good is *just* to say that they are possible good-*makers*. I am inclined to think that this is true, but I need not take a stand on this issue here. Perhaps what is a possible good-maker must also itself be good (and not just a good-maker) – what contributes to the goodness of things must itself have some goodness to contribute, as Dancy puts it.

(v) The possibility invoked in *Possible Value-Making* can be given different interpretations. Metaphysical possibility is one option. But some may claim that, on this interpretation, the account rules out too much, for isn't the state of affairs of *John's being a very happy cow* good even though it is metaphysically impossible that John is a very happy cow? Or, could we not say that the state of affairs of *there being a god who enjoys her creation* is good even if it is metaphysically impossible that there is a god? I think that we should not say that these states of affairs *are* good, only that they *would* have been good, had things been radically different (indeed, so different that essential truths would have been different). However, for my argument I only need weaker versions of the principle that invokes *logical* possibility, in the sense of what is compatible with the truths of logic. So understood, the *Possible Value-Making* will not rule out that the above-mentioned states of affairs are good, since it is metaphysically but not logically impossible that John is a happy cow and that there is a god.

If we accept this understanding of the value of abstract exemplifiable alternatives, we have an argument against the idea that existence can be better for you than non-existence. For all A, B, and S (all premises being necessarily true):

[29] As regards the *extrinsic* value of X, which is defined in terms of the final or intrinsic value of features other than X that would be exemplified if X were exemplified, it seems at least misleading to say that things can be good in virtue of (just) exemplifying X. Rather, here things can have value in virtue of exemplifying all the valuable features that were exemplified as a result of X being exemplified. To make room for this kind of value, we could simply weaken the principles further so that they say that if X is good (bad/neutral), then it is possible that something is good (bad/neutral) *and X is exemplified.* No matter whether X or its counterfactual dependents make things good, X itself must be exemplified. Similar considerations apply to states of affairs with *non-basic* intrinsic value. For example, suppose a certain exclusive disjunction of good states of affairs is itself good (e.g., your being happy to degree 5 or your being happy to degree 10), and that it is good in virtue of the good disjuncts, then one may want to claim that when this state of affairs is exemplified things are made good, not in virtue of the whole disjunction, but in virtue of the disjunct in virtue of which the disjunction is exemplified.

(1) If A is better (worse) for S than B, then the value of A for S is greater (less) than the value of B for S. (*Better-for entails value-for*)

(2) If A has a certain value for S, then it is possible that something has this value for S in virtue of exemplifying A. (*Possible Value-Making)*

(3) If something has a certain value for S in virtue of exemplifying A, then A is exemplified and S exists. (*Factivity*)

So

(4) If A has a certain value for S, then it is possible that A is exemplified and S exists. (from (2) and (3))

So

(5) If A is better (worse) for S than B, then it is possible that A is exemplified and S exists and it is possible that B is exemplified and S exists. (from (1) and (4))

So

(6) If the state of affairs of S's existence is better (worse) for S than the state of affairs of S's non-existence, then it is possible that S's existence is exemplified and S exists and it is possible that S's non-existence is exemplified and S exists. (from (5))

(7) It is not possible that S's non-existence is exemplified and S exists. (logical truth)

So

S's existence is not better (worse) for S than S's non-existence. (from (6) and (7))

Premise (1), *Better-for entails value-for*, states a very plausible principle. Indeed, something similar holds for all comparatives. If I am taller than you, then my height is greater than your height; if I am heavier than you, then my weight is greater than

your height. In general, if x is Fer than y (and 'Fer than' is a comparative), the Fness of x is greater than the Fness of y.[30]

Premise (2) is true since it is an application of the general principle *Possible Value-Making* to 'value for'. Remember that we are here talking about value-making by way of exemplification, which holds between abstract alternatives and their *exemplifiers*, not value-making by way of entailments, which holds between abstract alternatives when one valuable alternative is part of or entailed by another.

Premise (3), *Factivity*, is true, since 'value for S' entails the existence of S, and 'in virtue of' is factive: if x is F in virtue of x being G, then x is F and x is G and thus x exists.

Premise (7) is just a logical truth: necessarily, if S's non-existence is exemplified, then S does not exist.

Note that this argument, if sound, enables us to restate the challenge for PAC without relying on *Better for-Better off* (though I do believe that this principle is also true). Its conclusion together with PAC entails the problematic conclusion that in no non-identity case (of the mere addition form) can one alternative be better or worse than another.

9. Better-for does not entail better-off: guardian angels

At this point the opponents could say that even though some states of affairs, such as my non-existence, are not possible value makers, they can still merit certain *attitudinal responses* by existing people.[31] Drawing on this idea, Arrhenius & Rabinowicz present the following argument for the claim that a person's existence can be better for her than her non-existence.[32] We assume throughout that we are talking about final values for S and corresponding final preferences.

> (1) If A is good for S and S does not exist in B, then one ought to prefer A to B, for S's sake.

[30] That the Fness of x is greater than the Fness of y does not mean that Fness must always come in precise or determinate levels or degrees, for there can be some indeterminacy about Fness. In that case the Fness of something is better represented by a *range* of precise levels of Fness. In order to compare the Fness of x to that of y we would then need to compare the range of precise levels associated with x with the range associated to y. One simple idea would be to say that one range is greater than another if all precise levels in the former range is greater than those in the latter, but there are alternative ways of making sense of this comparison.

[31] Arrhenius & Rabinowicz (2013).

[32] A similar argument is presented in Bradley (2014).

(2) If one ought to prefer A to B, for S's sake, then A is better for S than B.

So,

(3) If A is good for S and S does not exist in B, then A is better for S than B.

Since S cannot be better off in A than in B, because S does not exist in B, the argument, if sound, also provides a counterexample to *Better-for entails Better-off*.

'for S's sake' can mean different things.[33] Sometimes

(T1) 'for S's sake' means 'insofar as one *respects* S'.

But then (2) is obviously false, since what is good or better for someone can come apart from what would show her most respect. It can be true that I ought to respect your decision to sacrifice your life for others, and thus that I ought to prefer that you do it, out of respect to you, even though not sacrificing yourself would make you better off.

Arrhenius & Rabinowicz seem inclined to think that 'for S's sake' means 'insofar as one cares about S'.[34] But then 'for the sake of S' can mean different things depending on *how* and *why* one cares about S. Is there a uniform interpretation of 'for S's sake' that makes the argument convincing?

(T2) 'for S's sake' means 'insofar as one cares about what is *better for* S'.

This reading would simply be question-begging, since, then, (1) on its own entails that 'If A is good for S and S does not exist in B, then A is better for S than B', for 'ought to prefer insofar as one cares about what is better for S' entails 'better for S'.

(T3) 'for S's sake' means 'insofar as one cares about what is *impartially* better in terms of S's wellbeing'.

I would agree that (1) is true, on this reading, since I think that A is in one respect impartially better than B in virtue of the fact that A is good for S and B is not. However, to insist that this shows that A is *better* for S than B, and thus to assume

[33] I am not suggesting that 'for S's sake of' shows the same ambiguity as 'bank'. It is more plausible to think it has a unified context-invariant meaning that determines what it expresses in a certain context. The standing meaning would then be something like 'on the grounds of some contextually salient purpose that involves S being favoured in some sense'.

[34] Arrhenius & Rabinowicz (2015), footnote 22, refers approvingly to a quote from Darwall (2002), in which 'should desire A for S's sake' is equated with 'should desire A insofar as one cares about S.'

that (2) is true, again, would be question-begging, since the issue is exactly whether A is better for S than the S-less B.

> (T4) 'for S's sake' means 'insofar one cares about what contains more wellbeing for S, or more value for S'.

This would make (1) true, since, as I pointed out in my discussion of Roberts' views, if A contains some wellbeing for S and B none, then, in one sense of 'more wellbeing', A contains more of S's wellbeing than B does, and it seems reasonable to say that one ought, for S's sake, prefer more wellbeing for S. But, again, it seems question-begging to assume that from the fact that there is more wellbeing for S in A than in B, just because there is some in A and *none* in B, it follows that A is better for S than B. As pointed out earlier, to say that there is no wellbeing for S in B is not the same as saying that B has neutral value for S.

None of these three interpretations gives us a non-question begging uniform reading of 'for S's sake' that would make both (1) and (2) clearly true. That the argument still may seem intuitively attractive can be explained by a shift in our understanding of 'for S's sake'. We find (1) true because we assume one interpretation and (2) true because we unwittingly assume another.

A last-ditch attempt to save the argument would be to just insist that there must be an *indefinable* sense of 'for sake of' that will fit the bill and make all premises true. But why should we believe that? We need some independent reason to believe in such a primitive sense of 'for the sake of'. Furthermore, even if we concede that there is such a primitive sense, (2) is still highly contestable. Here are some possible counterexamples that need to be addressed, at least if 'prefer' is understood as 'contemplate with greater pleasure (or some other positive contemplative positive attitude)':

> (a) One ought to prefer, for your sake, your eating disgusting mud to your not doing it, since the evil demon will kill you, if you do not eat the mud, but your eating disgusting mud is not (finally) better for you.[35]

> (b) One ought to prefer, for your sake, your life's being good for you to your life's being bad for you, but your life's being good for you is not better for you than your

[35] This is an example of the so-called 'wrong kind of reason'-objection, which has attracted a lot of attention recently. For a recent paper that has shaped much of this debate, see Rabinowicz & Rønnow-Rasmussen (2004).

life's being bad for you. The value for you of your life's being good for you is not greater than the value for you of your life's being bad for you, since if *evaluative* states of affairs, such as your life's being good for you, themselves have value for you, an infinite regress of values for you looms: A has value for you, (A's having value) has value for you, ((A's having value) having value for you) has value for you...

(c) One ought to prefer, for your sake, your being a happy cow to your being an unhappy cow, even though your being a happy cow is not better for you, since it is *metaphysically impossible* for you to be a cow (assuming you are a human). Of course, if you *had* been a cow, then it *would* have been better for you to be a happy cow.

(d) One ought to prefer, for your sake, your being extremely happy and the only barber who shaves all and only those who do not shave themselves to your being unhappy, but the former state of affairs is not better for you since it is *logically impossible*.

10. Concluding remarks

None of the attempts to ease the tension between PAC and our considered judgements seems promising. A more plausible way to avoid the counterintuitive conclusion that no outcome is better than another in non-identity cases is simply to abandon PAC. I would like to end on a more conciliatory note, however. Arrhenius & Rabinowicz and I agree that a person is better off (or worse off) in one alternative than in another only if she exists in both alternatives. So, we all agree that a person cannot be affected for better or worse by being created. And we also think that an alternative can be better without anyone being better off. Unlike Fleurbaey, Voorhoeve and Roberts, Arrhenius & Rabinowicz agree with me that this principle should be rejected:

> If A is better than B, then some person (in A or in B) is better off in A than in B.

So, it is clear then that neither Arrhenius & Rabinowicz, nor I, accept a *strict* person-affecting constraint, according to which what is better must have *affected* some individual. Instead, we accept:

Extended person-affecting principle (EPAC):

If A is better than B, then

(a) some person (who exists in both A and B) is better off in A than in B, or

(b) A is good for a person who exists in A but not in B, or

(c) B is bad for a person who exists in B but not A.[36]

This is still a person-affecting constraint, but in less strict sense: what is better must make a positive difference in individual well-being facts. After all, to create someone who will be well off is also a way of affecting individual wellbeing, since it is a way to make a positive difference in individual wellbeing facts, even though it does not make anyone better off. It is positive difference in wellbeing facts, since what is added is something that is good for people.[37] Similarly, to create someone who will be badly off is also a way of making a negative difference in individual wellbeing facts, even though it does not to make anyone worse off.[38] It is negative, since what is added is something that is bad for people.[39]

The difference between Arrhenius & Rabinowicz and me is that they want to add that in case (b) it is true in A, where S exists, that A is better for S than B, and that in case (c) it is true in B, where S exists, that A is better for S than B. I strongly doubt that it is really worth violating reasonable principles for value-for, such as *Possible Value-Making,* just to be able to maintain a *comparative* form of a person-affecting constraint - 'better' entails '*better* for someone', especially when it is already agreed that non-identity cases show that an alternative can be better without affecting anyone for the better.

[36] More precisely: If A is better than B, then
(a) either if A were the case, then it would be the case that someone is better off in A than in B, (or, equivalently, if B were the case, then it would be the case that someone is better off in A than in B),
(b) If A were the case, then it would be the case that A is good for someone who would not exist if B were the case,
(c) If B were the case, then it would be the case that B is bad for someone who would not exist, if A were the case.
[37] This could be seen as a *non-comparative* benefit, as McMahan (2013) has recently argued.
[38] For more on different forms of person-affecting restrictions, see Arrhenius (2009).
[39] This could be seen as a *non-comparative* harm, as McMahan (2013) has recently argued.

References

Adler, M. D. Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis, Oxford University Press, 2011.

Adler, M. D. "Future Generations: A Prioritarian View", The George Washington Law Review, Vol. 77, No. 5/6, September 2009.

Arrhenius, G. "Can the Person Affecting Restriction Solve the Problems in Population Ethics?" in M. A. Roberts and D. Wasserman (eds.), Harming Future Persons. Aldershot: Ashgate, 2009, pp. 289–314.

Arrhenius, G. "The Affirmative Answer to the Existential Question and the Person Affecting Restriction." in I. Hirose and A. Reisner (eds.), Weighing and Reasoning, Oxford University Press, 2014.

Arrhenius, G. and W. Rabinowicz. "Better to Be Than Not to Be?" in H. Joas and B. Klein (eds.), The Benefit of Broad Horizons. Intellectual and Institutional Preconditions for a Global Social Science, Festschrift for Björn Wittrock, Brill: Leiden, 2010, 399–421.

Arrhenius, G. and Rabinowicz, W. 'The Value of Existence' in Hirose, I. and J. Olson (eds.), Oxford Handbook of Value Theory, 2015.

Bradley, B. 'Asymmetries in benefitting, harming, and creating', Journal of Ethics, 2014.

Bradley, B. Wellbeing and Death, Oxford: Oxford University Press, 2009.

Broome, J. Ethics out of Economics, Cambridge: Cambridge University Press, 1999.

Broome, J. Weighing Lives, Oxford: Oxford University Press, 2004.

Carlson, E. "'Good' in terms of 'Better"", Noûs 50:1, 2016, pp. 213–223.

Chisholm, R. M. and Sosa, E. "On the Logic of 'Intrinsically Better'", American Philosophical Quarterly, 3, 1966, pp. 244-249.

Darwall, S. Welfare and Rational Care. Princeton, NJ: Princeton University Press, 2002.

Fleurbaey, M. and A. Voorhoeve, "On the Social and Personal Value of Existence." In I. Hirose and A. Reisner (eds.), Weighing and Reasoning: A Festschrift for John Broome. Oxford University Press. 2014.

Glover, J. (1977). Causing Death and Saving Lives. New York: Penguin.

Gustafsson, J. E. "Neither 'Good' in Terms of 'Better' nor 'Better' in Terms of 'Good'", Noûs, online prepublication, DOI: 10.1111/nous.12038, 2013.

Gustafsson, J. E. 'Still Not "Good" in Terms of "Better"' Noûs, 50 (4), 2016, pp. 854–864.

Hare, R. "Possible People." Bioethics 2, 1988, pp. 279–93.

Hirose, I. and Reisner, A. (eds.) (2015) Weighing and Reasoning: Themes from the Philosophy of John Broome, Oxford: Oxford University Press.

Holtug, N. "Person-Affecting Moralities." In J. Ryberg and T. Tännsjö (eds.), The Repugnant Conclusion: Essays on Population Ethics. Dordrecht: Kluwer, 2004, pp. 129–61.

Holtug, N. "In Defence of the Slogan." In W. Rabinowicz (ed.), Preference and Value: Preferentialism in Ethics. Studies in Philosophy, vol. 1. Lund: Department of Philosophy, Lund University, 1996, 64–89.

Holtug, N. "On the Value of Coming into Existence." Journal of Ethics 5, 2001, pp. 361–84.

Johansson, J. "Being and Betterness." Utilitas 22, 2010, pp. 285–302.

Lewis, D. On the Plurality of Worlds, Oxford: Blackwell Publishers, 1986.

McMahan, J. 'Causing people to exist and saving people's lives', Journal of Ethics 17, 2013.

Narveson, J. "Utilitarianism and New Generations." Mind 76 (January), 1967, pp. 62–72.

Parfit, D. Reasons and Persons. Oxford: Clarendon Press, 1995 (1984).

Rabinowicz W. and Rønnow-Rasmussen T. 'The Strike of The Demon. On Fitting Pro-Attitudes and Value', Ethics 114, April, 2004, pp. 391–423.

Roberts, M. A. Child versus Childmaker: Future Persons and Present Duties in Ethics and the Law. Lanham, MD: Rowman and Littlefield, 1998.

Roberts, M. A. "Can It Ever Be Better Never to Have Existed at All? Person-Based Consequentialism and a New Repugnant Conclusion." Journal of Applied Philosophy 20 (2), 2003, pp. 159–85.

Temkin, L. S. Inequality. Oxford: Oxford University Press, 1993a.

Temkin, L. S. “Harmful Goods, Harmless Bads.” In R. G. Frey and C. W. Morris (eds.), Value, Welfare, and Morality. Cambridge: Cambridge University Press, 291–324. 1993b.

Williamson, T. Modal Logic as Metaphysics, Oxford University Press, 2013.

Zimmerman, M. J. The Nature of Intrinsic Value, Lanham, Md.: Rowman & Littlefield, 2001.

M.A. Roberts[1]

What Is the Right Way to Make a Wrong a Right?

It seems clear that the most challenging versions of the nonidentity problem involve, at least implicitly, claims about probability. Once we realize that, we are tempted to appeal to the concept of *expected utility* for purposes of understanding the problem and analyzing the underlying cases. But there are reasons to think that that approach is ultimately unsatisfactory. Thus the question remains open just how probabilities are to be brought to bear in connection with nonidentity. This paper explores some of our options and some of the challenges those options will face.

[1] Philosophy, Religion, and Classical Studies Department, The College of New Jersey, robertsm@tcnj.edu.

1. Trading Off the Better Outcome Against the Better Chance

It is plausible to think that not all choices that end badly—end, for example, in outcomes in which a particular person has less wellbeing than that same person could have had at no cost to anyone else—are wrong. Consider the case of *Fertility* (Graph 1).

Graph 1. Fertility

		0.0001	0.0099	0.99
c1 potential mom doesn't undergo fertility treatment	 +10 +8 +0	*o1* **p** *p*	*o2* *p*	*o3* *p*
c2 potential mom undergoes fertility treatment	 +10 +8 +0	*o3* *p*	*o4* **p** 	*o6* *p*

In this graph and the following, the acts, or choices, c1, c2 and so on exhaust the agent's (agents') available choices and the possible worlds, futures or outcomes, o1, o2 and so on exhaust the outcomes accessible under those choices.[2] *The numbers 0-1 in the second row represent the probabilities, based on information available to the agent(s) just prior to choice, that a given outcome will unfold under a given choice. The numbers in the second column represent (overall, lifetime; raw, unadjusted) wellbeing levels. (It's left open whether wellbeing consists of happiness, capability or something else entirely.) Positive numbers represent a life worth living and negative numbers a life less than worth living. p, q and so on are constants reserved for persons, with the term person being understood to include many non-human animals and not all humans. A personal constant in bold means that the person does or will exist, and in italics that the person never exists, in the particular outcome. It is assumed that a person's wellbeing level is zero in any outcome in which that person never exists (no wellbeing burdens; no wellbeing benefits) and stipulated that no one other than the persons identified in the graph is affected (in terms of existence, wellbeing or any other potentially morally relevant respect) however the choice under scrutiny is made. A gray background indicates the choice that is in fact made in the particular case and the outcome that in fact unfolds under that choice.*

[2] I use the term "*accessibly* could have had," rather than simply "*could* have had," in recognition of the fact that not all logically possible outcomes, however wonderful, are outcomes morality obligates agents to bring about. The fact that it's logically possibly that one choice leads to a better outcome than another choice doesn't, we think, imply that the other choice is wrong in the case where the better outcome is barred by, e.g., the laws of physics.

In Fertility, a woman, the agent, must choose between undergoing an effective and relatively safe treatment for her infertility or declining that treatment. The woman in fact chooses treatment—she chooses c2 over c1—and outcome o4 in fact unfolds and the child p in fact comes into existence. But just *prior* to choice the not-yet-existing p faces an across-the-board profound existential risk: c2 makes p's existence *more* likely than c1 does, but neither c1 nor c2 *assures* p's future existence or even makes that future existence at all *likely*. All that is likely and is indeed *assured* is that, if p does come into existence—if, that is, o1 or o4 obtains—then p will have an overall, lifetime wellbeing level of +10 under c1 or +8 under c2. c2 thus *caps* p's wellbeing level at +8; under c1 and c1 alone is there *any* chance that p's wellbeing will be *maximized* at +10.

But we don't think c2 is *morally wrong*. Rather, we think c2 is perfectly *permissible*. It's true that *but for* the probabilities—the probability being *relatively* high (though not *high*) that p will exist under c2, that is, that o4 will unfold given c2, and the probability being relatively low (indeed, *low*) that p will exist under c1, that is, that o1 will unfold given c1—we *would* want to say that c2 is wrong and that c1 is obligatory. The probabilities being as they are, however, we consider c2 is permissible.

In this case and many others, the probabilities have served to convert an otherwise wrong choice into a permissible choice; they have converted a *wrong into a right*. We can surmise that the applicable moral principles generating the result of permissibility function to *trade off* a better *outcome* for a given person against a better *chance* of that person ever existing at all. Because p has a better *chance* of existing under c2, we say that c2 is permissible even though p's *wellbeing* level in o4 is less than what p *accessibly could have had* (at no cost to anyone else) in o1 under c1.[3]

To show that the distribution between wellbeing and chance we see in Fertility can arise in real life, we need just add some details to the case. Suppose that the woman is suffering from an elevated prolactin level, a condition that reduces the woman's chances of ovulation and thus her chances of conceiving a child. To lower her prolactin level, she has the option of taking the drug bromocriptine. But drugs often come with unwanted side effects, and let's now just stipulate (the actual medical reports on the side effects of bromocriptine being mixed) that bromocriptine leaves any child it enables a woman to conceive with adverse skin and neurological conditions. But let's also stipulate that those side effects are fairly mild;

[3] I use the term "*accessibly* could have had," rather than simply "*could* have had," in recognition of the fact that not all *logically* possible outcomes, however wonderful, are outcomes morality obligates agents to bring about. The fact that it's *logically* possibly that one choice leads to a better outcome than another choice doesn't, we think, imply that the other choice is *wrong* in the case where the better outcome is *barred* by, e.g., the laws of physics.

that the child's existence will remain clearly worth having; and that no other treatment nearly as effective as bromocriptine is available.

Those details in place, we, perhaps even more clearly than before, see *nothing wrong* with the woman's choice of c2.

Of course, *prior to choice*, there's no *one* child—p—whom the woman will be assured of conceiving if she conceives at all—it may be p, or q, or r, or any one of many other possible children depending on which sperm accomplishes the actual insemination. After all, there may be 200 million or more sperm cells in a single human ejaculate. (One aside: that biological fact means that the probabilities in Fertility are wildly, upwardly, exaggerated, in order to make the arithmetic we will get to a little later a little more intuitive.) Thus p's chances—indeed any *particular* future person's chances in *any* case—of coming into existence will typically remain low whatever the agent does. But the distribution between wellbeing and chance we see in Fertility applies not just to p but also to *any* child the woman might conceive under the scenario described. And for *any* such child we will think the same thing: that the probabilities at stake in the particular case have converted an otherwise wrong choice into a clearly permissible choice.[4]

2. Expected Value

A widely accepted way of accounting for the intuition of permissibility in cases like Fertility makes use of the concept of *expected wellbeing*, that is, *expected value* (*EV*).[5] Here, by stipulation, it's just the plight of the *one child p* that is at stake; it's how *that* child is, or is expected to be, affected that we need to focus on for purposes here. Now, I think that's how we *should* think about these things; I think, that is, that a person-based approach, rather than an impersonal approach, seems plausible and that we *should* take into account the plights of one person at a time rather than the mass of all persons in aggregate and rather than the *universe* per se. Accordingly, I've constructed the expected value calculation so that we can do just that.[6]

[4] Prior to choice, the woman—the agent—cannot, of course, refer to any of her potential children by a genuine proper name. But her ability to quantify over classes of individuals for whom she has no such names means that she can nonetheless proceed with an ex ante moral evaluation of her available choices.

[5] For a clear introduction to expected value theory, see Feldman 2006. In that paper, Feldman argues against expected value theory on grounds of impracticality—grounds other than those I describe in this paper.

[6] But the formula at work here in the expected value calculation is itself perfectly standard. Thus, if, in the end, we decide *against* thinking about things in that way—one person at a time—we can easily move from what I have written here to an impersonal formula if that's what we want to do.

> The *expected value* of a given choice c for a given person i (EV(c, i)) =
>
> > the summation of the values for i of the outcomes accessible[7] under c multiplied by the respective probabilities that those outcomes will obtain given c.

We can then calculate as follows:

> EV(c1, p) = 10 x .0001 + 0 x .0099 + 0 x .99 = .001; and
>
> EV(c2, p) = 0 x .0001 + 8 x .0099 + 0 x .99 = .0792.

Since EV(c2, p) > EV(c1, p), c2 is the choice that maximizes expected value for p.

It may well seem that—and has indeed in the past seemed to me that—on that basis alone we can then declare c2 permissible. Plausibly, the principle that would generate that conclusion for us would be a *simple necessary* condition on wrongdoing, one that says that, when other things are equal as they are in Fertility, a choice that maximizes expected value for p is permissible. Such a principle, of course, leaves open the question whether c1 is permissible as well. But it also—plausibly in my view; more on this later—avoids the implication that c1 is wrong.

Now, aiming to develop a cogent, plausible, consistent person-based approach, I would add to that simple necessary condition an *existence* condition. The idea would be to make it clear that the principle *doesn't* open the door to the claim (and of course, as a mere *necessary* condition on wrongdoing, wouldn't in any case *imply*) that a given choice is wrong in virtue of the fact that it has failed to maximize expected value on behalf of a person in an outcome in which that person never exists at all.

Putting those two conditions together, we obtain the *expected value maximizing principle* (*EVMP*).

> *Expected value maximizing principle (EVMP).* A choice c made in a given outcome x is wrong *only if*

[7] Here we could just say "possible" rather than "accessible" since any outcome inaccessible under c will be one the probability of which is zero given c. The reverse may not hold: it may be that the probability of a given outcome under a given choice is zero even in a case in which the outcome itself is accessible.

there is a person i who does or will exist in x under c and

there is an alternate available choice c′ such that the expected value of c for i is less than the expected value of c′ for i.

Under EVMP, c2, having *maximized* expected value for p, *isn't* wrong—is, that is, *permissible*. And, for the case of Fertility, that's, of course, the result we want.

Now, some expected value theorists may want to say more—more, that is, than what a simple *necessary* condition on wrongdoing can imply. They may push to say that c2 is not only permissible but also *obligatory*—that is, that c1 is *wrong*. In what follows, I will return to that issue, and spell out why I think we may not want to take that further step (part VI below). But our focus, for now, is on what it is about the probabilities that makes c2 *permissible*. So for now we set aside the issue of whether c1 is wrong.

3. The Nonidentity Problem

The most challenging nonidentity cases are the ones in which the choice under scrutiny is clearly wrong. Those cases include: the choice of the risky over the safe environmental policy (Parfit 1987); the choice to deplete rather than conserve resources (Parfit 1987); the choice to do nothing rather than something about climate change (Broome 1992); choices yielding historical injustices; and the choice to sign the slave child contract, or to take the iatrogenic pleasure pill, just prior to conceiving a child (Kavka 1981).

Now, it might *seem* that those cases display the same tradeoff between outcome and chance that we see in Fertility. If that's so, and if the argument to permissibility is valid in Fertility, it must be valid in those nonidentity cases as well. And that in turn would mean that, though we *like* the result that EVMP generates in the case of Fertility, we would be compelled nonetheless to reject EVMP on the ground that it generates clearly false results in the nonidentity cases.

But that wouldn't mean that we should abandon the idea that *expected value* has a critical role to play. The defect might instead lie in the EVMP's *existence condition*. We could get rid of that and ask, not whether expected value has been maximized for a *particular* existing or future child p, but rather whether there's *any* child q for whom an alternate choice would have produced still more expected value. If so, then the expected value principle, stripped of the existence condition, would avoid the result that the choice under scrutiny—risky policy, depletion, pleasure pill, etc.—is permissible.

Do we thus have, in the nonidentity cases a compelling objection against EVMP —and, specifically, against the existence condition?

No. The objection relies on the assumption that the pattern of tradeoffs between the better outcome and the better chance in the nonidentity cases is just what we see in Fertility. But that assumption is false.

In Fertility, the woman chooses to undergo a treatment that *increases her fertility* and thus, for her child p who in fact exists, made it more likely, *calculated at that time just prior to choice*, that p, in due course, would exist. It's just an actual outcome *bias*, an unexamined and faulty *assumption*, an instance of *post hoc ergo propter hoc*, that the choices under scrutiny in the nonidentity cases will do that same work: that choosing depletion over conservation—or choosing the risky over the safe policy, or choosing to do nothing rather than something about climate change, or choosing to take the pleasure pill or sign the slave child contract—will similarly make whoever it is who ultimately exists and suffers as a result of those choices more likely, *calculated at that time just prior to choice*, to exist (Roberts 2007; 2009; Roberts and Wasserman 2016).

And it's an assumption that badly fails when exposed to the clear light of day. The child who is conceived just after the would-be parent takes the pleasure pill is no more likely to exist under that choice than that *same* child is to exist under the choice to take the aspirin instead.[8]

So: so far so good for EVMP, existence condition in place. It generates the correct result for the case of Fertility and avoids clearly incorrect result in the most challenging versions of the nonidentity problem.

4. Two Problem Cases for Expected Value

But we do have, it seems, independent reasons to question both the person-based and the impersonal appeals to expected value.

Consider a second mother and child case, the case of the *All-But-Known Disaster* (Graph 2).

[8] And it's not just that we can't look into the future and *identify* any such person and say, of that person, that his or her chances of coming into existence are improved. That practical problem is readily resolved in virtue of the fact that we can use quantifiers to talk about those future people. For no such person i who might exist following the parent's ingestion of the pleasure pill is i's chances of existence greater under the parent's choice to take the pleasure pill than it is under any alternate, safer choice (e.g., the choice to take the aspirin, or sip of water, in place of the pleasure pill).

Graph 2. All-But-Known Disaster[9]

		.9999	.0001
c1		*o1*	*o2*
mom doesn't choose	+1M		
risky extension	+8	**p**	**p**
on behalf of child p			
	+0		
EV(p, c1) = 8 x .9999 + 8 x .0001 = 8	-10		
c2		*o3*	*o4*
mom chooses	+1M		**p**
risky extension	+8		
on behalf of child p			
	+0		
EV(p, c2) = -10 x .9999 + 1M x .0001 = 90.001	-10	**p**	

M = 1,000,000.

Two important features of this case should be highlighted. First, in this new case, the child p faces *no* existential risk. Perhaps p already exists or that p will exist whatever choice is made. Second, c2 is a single, one-off, high risk/high reward, choice. To be clear: the option of the mother's, the agent's, pursuing a *long-term* strategy that allows for the choice of c2 *in the context of* a sequence of choices that includes not just c2 but also successive c2-like choices, choices that repeat the risks and rewards generated by c2, is *not* a choice available to the agent. Thus, p may end up—and in the particular case *does* end up—as the *single trial non-beneficiary* of the high risk/high reward c2.

To give some color to the case, we can suppose that a mother faces the choice whether to have her only child, p, undergo a very risky but possibly enormously beneficial treatment, a treatment that might extend p's life for thousands of years. If she declines treatment—chooses, that is, c1—then p is certain to have, at +8, a good life however the future unfolds. Specifically, let's imagine that at +8 p's wellbeing level is about what the woman herself, as well as what others in her and her child's generational cohorts, will enjoy. In contrast, if she chooses treatment—chooses, that is, c2—then the probability is *very high* that p will have, at -10, an existence that is significantly *less* than an existence worth having. In fact, she *all but knows* that, if she chooses c2, p will be thoroughly miserable; she *all but knows* that c2 will be a complete disaster for her child. And it is.

Is her choice of c2 wrong? We would easily, unobstructedly, say that it is wrong, *but for one final detail of the case*: the very small probability that the choice of treat-

[9] I am grateful to Dean Spears for the case on which All-But-Known Disaster is based.

ment will unfold into an outcome that benefits p enormously—that generates for p an extraordinarily good life, say, a life that is *both* very good *and* goes on and on for thousands of years.

Here, we calculate as follows:

EV(c1, p) = 8 x .9999 + 8 x .0001 = 7.9992 +.0008 = 8; and

EV(c2, p) = -10 x .9999 + 1M x .0001 = -9.999 +100 = 90.001.

If the concept of expected value is the right way to bring probability to bear in moral analysis—if a choice that maximizes expected value for one person and leaves everyone else alone can't be wrong—then we should conclude that c2 is permissible.[10]

But can the result that c2 is permissible be correct? Do we think, in the case of All-But-Known Disaster, that the fact that p might—*might*; *against all odds,* the chances of things turning out well for p under c2 being a scant 1 in 10,000—obtain an enormous benefit converts what would otherwise be a clearly *wrong* choice into a *permissible* choice?

I don't think that we do. The concept of expected value helped to make sense of our intuitions in Fertility *and* the nonidentity cases. But it *doesn't* help us make sense of All-But-Known Disaster *at all*. Though c2 clearly maximizes *expected value* for the one existing or future person p who might possibly be affected by how the choice is made, our clear intuition is that c2 is *wrong*.

To test that intuition further—to make the case more vivid, but not to change facts of the case or in the end the moral analysis—let's just note that it's part of the case that the woman *in fact* chooses c2 and that the predicted outcome, o3, the outcome in which disaster strikes, *in fact* obtains. We see the miserable child; we see the hapless mother. It now seems clearer than ever that c2 is wrong.

Of course, in *another case altogether*—call it *Long Run Risk Reduction*—we may well say that c2 is permissible. In that case, the woman has the option of completing a *sequence* of high risk/high reward choices, commencing with c2 and continuing with (perhaps thousands) of successive c2-like choices, choices that repeat the risks and rewards for p that come with c2 itself. *Provided* that the available sequence is sufficiently long *and* that the mother in fact plays out the long-run strategy—provided, that is, that c2 is chosen in (relative to) an outcome that meets those con-

[10] Indeed at least some theorists want to go beyond the result that c2 is permissible and say more than a simple necessary condition on wrongdoing can imply—that is, that c1 is *wrong*. But again for the moment we'll set that issue aside and focus exclusively on the question of c2's permissibility.

ditions—c2 plausibly *is* permissible. For then the long term strategy, itself commencing with c2, will come with a significantly high probability that *at some point* p will luck out and accrue an enormous benefit, thereby, at least over the long run, significantly reducing the risk of disaster for p.

But All-But-Known Disaster is *not* that sort of case. In All-But-Known Disaster, the accessible outcomes that include the choice of c2 are limited to just o3 and o4. And in each of those outcomes it's stipulated that c2 exists as just a *single, lonely, one-off* choice and *not* as the start of a risk-reducing sequence of further c2-like choices.[11]

One other quick note. Some theorists might bring the concept of risk aversion into play at this point to explain away the intuition that c2 is clearly wrong, or perhaps to allow us to say that, relative to the risk averse agent, c2 is wrong, but relative to the risk neutral agent, c2 is permissible. And it's not implausible to suppose that those claims can form the start of a plausible account of Long Run Risk Reduction.

But All-But-Known Disaster is a very different sort of case. In that case, it seems highly implausible that even the evaluator *neutral* in respect of risk will consider c2 permissible. Even the *neutral* evaluator will agree that it's wrong for the woman, when she *all but knows* c2 will end in disaster for her child, to choose c2. Considerations of risk aversion thus seem beside the point in All-But-Known Disaster.

Here is a second problem case for the expected value approach, one that does involve existential risk.

[11] In general, it's a mistake to think that what we say about the permissibility of a choice (an act) made (performed) in one outcome in one case must always be what we say about the permissibility of that same choice (act) made (performed) in another outcome in another case. Thus the fact that c2 is permissible relative to (when chosen in) certain outcomes available to the agent in Long Run Risk Reduction gives us *no reason at all* to say that c2 isn't wrong relative to (when chosen in) o3 in All-But-Known-Disaster. At least, that that would be a mistake is an essential tenet of any consequentialist approach. As Feldman points out, the same medicine might either save or kill a patient, depending on whether or not it's followed up by a second medicine, and we can consistently take the position that giving the one medicine is permissible relative to (when made in) the outcome in which it's followed up by the second but impermissible relative to the outcome in which it's not. Feldman 1986.

Graph 3. 50-50 Disaster[12]

		.5	.5
c1 agent chooses to bring child into existence under high risk/high reward condition	 +101 +0 -100	*o1* **p**	*o2* **p**
c2 agent chooses not to bring child into existence at all	 +100 +0 -100	*o3* *p*	*o4* *p*

The potential upside for p being just a bit greater than the potential downside, EV(c1, p) > EV(c2, p). According to EVMP, c1 is permissible. But that seems implausible. Where the agent has the option of not bringing the child into existence to begin with—the option to assure that the child will *not* be forced to endure a deeply, genuinely *wrongful* life—isn't the only permissible choice the choice not to bring the child into existence to begin with?

* * *

Thus the appeal to expected value turns out to be problematic. Yet we can't let go of the idea that probabilities play a critical role in moral evaluation. Fertility and many other cases as well tell us that that's so. My own proposed solution to the nonidentity problem in its most challenging forms, moreover, depends critically on that being the case.

The question thus arises whether probabilities can be brought to bear in the analysis by a route other than that of expected value.

5. Probable Value

One such possible route lies in the concept of *probable value* (*PV*). The underlying idea is that what makes the difference between the moral status of c1 and c2 in All-

[12] I am grateful to Mark Budolfson for this case. Risk and Population Workshop, University of Texas, Austin, Texas (Nov. 22-24, 2019).

But-Known Disaster is a substantially narrower set of facts than what we are required to take into account in calculating expected value. The EV calculation takes *all outliers* into account. In All-But-Known Disaster, the tiny chance of the extraordinary o4 obtaining under c2 forced us to say that c2 was permissible.

Under a probable value approach, the tiny chance of o4 obtaining would not be relevant to the mechanics of bringing probability to bear in the context of moral evaluation. Rather, under a probable value approach, where c2 ends in disaster, as in o3, what is critical in evaluating c2 at o3 is just the high probability, calculated just prior to choice, that o3 would in fact unfold under c2. The very high probability that o3 will obtain given c2 is the probability-significant feature that we are to refer to when evaluating c2 relative to o3. It's *that* number that we should multiply by the value that o3 has for p, to produce the *probable value of o3 for p under c2* (PV(o3, p, c2)). And the necessary condition on wrongdoing that we should adopt is the failure to maximize not *expected* value but rather the failure to maximize *probable* value.

The definition of probable values comes in two parts:

> *Probable Value.*
>
> > Where a choice c made in (relative to) an outcome x creates a probability n (calculated just prior to choice on the basis of information available to agents at that time) that a person i will have the wellbeing level (wb) that i in fact has at x, the *probable value* of x for i under c = the wellbeing i has at x multiplied by n.
> >
> > Where the *minimal wellbeing level* (*mwb*) for a person i at the range r of outcomes that may accessibly obtain under c is the *least* wellbeing level i has in any member of r, and where that choice c creates the probability n (calculated just prior to choice on the basis of information available to agents at that time) that some member or another of r will obtain, the *probable value* at r for i under c = n(mwb).

That is: PV(x, i, c) = n(wb of i at x); and, where the probabilities of various outcomes under a given alternate choice are each very low but the alternate choice nonetheless *dominates* (we can say) the one choice, PV(r, i, c) = n(mwb of i at r).[13]

In All-But-Known Disaster, then, the probability given c2 that p will end up at a wellbeing level of -10 is .9999, making PV(o3, p, c2) = -9.999. In contrast, we see

[13] I am grateful to Tomi Francis for a case that demonstrates the need for the second part of the definition of probable value. Nonidentity Workshop, Institute for Future Studies, Stockholm (Feb. 8-9, 2020).

higher probable values for both c1 at o1 and c1 at o2: PV(o1, p, c1) = 8 x .9999; and PV(o2, p, c1) = 8 x .0001.

The simple necessary condition, then, on wrongdoing would be that a wrong choice can't be one that maximizes probable value for each and every existing or possible person. For the person-based take on things, we would, as before, add the existence condition. The result is the *probable value maximizing principle* (*PVMP*):

> *Probable value maximizing principle (PVMP).* A choice c made at (relative to) a given outcome x is wrong *only if*
>
> there is an individual person i who does or will exist in x and there is
>
> (i) an alternate available choice c′ made at (relative to) an alternate accessible outcome y such that PV(x, i, c) < PV(y, i, c′) or
>
> (ii) z is a member of the range r of outcomes that may accessibly obtain under c′ and PV(x, i, c) < PV(r, i, c′).

There are, then, three ways of satisfying the proposed probability-related necessary condition on wrongdoing: one is for there to be an alternate choice at an alternative outcome that creates more probable value; another is for there to be an alternate choice and an alternate range of outcomes that shows that the alternate choice dominates the one choice; and still another is for the person for whom things look to be mathematically unfortunate never to have existed at all in the particular outcome at all.

Since PV(o3, p, c2) *isn't* maximized—it's *less* than PV(o1, p, c1) and *less* than PV(o2, p, c1)—nor is it dominated by any alternate choice, PVMP nicely avoids the false result that c2 at o3 is permissible.

Thus PVMP—in contrast to EVMP—leaves the door open for us to say (under still other principles) that c2 at o3 is wrong. *That is progress.*

6. Actual Value Condition on Wrongdoing

Let's go back to Fertility. I think we can mine that case for still a third necessary condition on wrongdoing. Consider c1 at o1. There, p, against all odds, has made it into existence even though the woman *hasn't* chosen to undergo the fertility treatment, and p's maximized wellbeing has come at no cost to anyone else. It seems plausible to me to say that c1 made at (relative to) o1 is permissible.

The principle that says just that is the *actual value maximizing principle* (*AVMP*).

> *Actual value maximizing principle (AVMP)*: A choice made at (relative to) a given outcome is wrong *only if* there is a person who does or will exist in that outcome and an alternate available choice and an alternate accessible outcome such that the actual value of the one outcome for that person under the one choice is less than the actual value of the alternate outcome for that person under the alternate choice.

In other words: AV maximization for a given person, other things equal, implies permissibility.

7. Note on the Existence Condition

The existence condition, as things stand, has been built into both PVMP and AVMP. But I want to underline that the condition *isn't* that the person exist under *both* choices or in *both* outcomes. The requirement is just that a person for whom things are worse does or will exist in the one outcome.

In other words: that a person never exists at all, other things equal, implies permissibility.

The Narvesonian idea behind the existence condition is that the fact that a person never exists in a given outcome under a given choice doesn't, morally, count against that outcome or against that choice: that a person *doesn't* exist doesn't, other things equal, make the outcome worse or an otherwise permissible choice wrong. Under that principle—the existence condition—we find in Fertility that c2 at o3 is perfectly permissible. For the same reason, so is c1 at o2.

8. Consolidated Principle

We can consolidate the three various necessary conditions noted above into a single principle, one that makes use of the concept of probable value in explaining how probabilities are to be brought to bear in our evaluation of choices; that accepts that a person's *nonexistence* doesn't (other things equal) count either against the outcome or against the choice; and that accepts that, when we (whether against all odds or not) *succeed* in maximizing wellbeing for a given person, what we have done isn't, in the end, wrong (though it may well have *looked* wrong starting out).

I should note that the combined principle does more than just sum up the three necessary conditions. It also takes the substantive position that variations in proba-

bility bear solely on the evaluation of the agent's alternative *choices* and not at all on the ranking of *outcomes* in respect of their overall betterness. That means that, while embedding in an outcome a probability of—for example—.0099 in place of a probability of .0001 may be critical to the evaluation of the choices made within that outcome, it *isn't* critical to the ranking of the one outcome against still other accessible outcomes; it can't, on its own, make one outcome better or worse than another. What makes one *outcome* better or worse than another is a matter confined to how wellbeing is distributed across a given population and what alternate outcomes are accessible in the particular case. What makes one *choice* permissible rather than wrong may turn on the probabilities.

Thus the combined principle must be spelled out in two parts, the telic (relevant to the ranking of outcomes) and the deontic (relevant to the evaluation of choice).

PVMP+AVMP+EC:

Telic component: Where y is accessible relative to x, x is worse than y, only if there is a person i and an alternate outcome z accessible relative to x such that:

i does or will exist in x; and

x is worse for i than z is (where z may, but not need, be identical to y).

Deontic component: A choice c performed at x is wrong, only if there is a person i, an alternate choice c' and an alternate accessible outcome y accessible relative to x such that:

i does or will exist in x;

x is worse for i than y; and either

PV(x, i, c) < PV(y, i, c') or y is a member of the range of outcomes that may accessibly obtain under c' and PV(x, i, c) < PV(r, i c').

On the choice side, this principle provides a complete evaluation of Fertility: all choices at all outcomes are permissible (yay!) and a nearly complete evaluation of All-But-Known Disaster: c1 at o1 is permissible, and so is c2 at o4. But what does it say about the nonidentity problem?

9. The Nonidentity Problem Revisited

Earlier, I claimed that EVMP opens the door to a plausible account of the most challenging nonidentity cases.

In this part, we turn to whether PVMP provides an equally plausible account of the nonidentity cases.

And I'll just note that it easily does. Let's, just to keep things simple, focus on the pleasure pill case. Since the burdened child p's chances of existence are just as great whether the parent takes the pleasure pill or—say—an aspirin, the choice to take the pleasure pill relative to the outcome that in fact unfolds—the outcome in which the child in fact exists and suffers from the pill's side effects—fails to maximize PV. Nor does the choice maximize AV. And, finally, we simply note that the burdened child in fact exists in the particular outcome and under the particular choice under scrutiny.

That means that all three necessary conditions identified in the consolidated principle are satisfied. That fact, in turn, opens the door for still other principles—other person-based principles, I would surmise—to step in and say that the choice that the parent has made at that outcome is wrong.

To put the point another way: the concept of probable value positions us to provide a plausible account of the pleasure pill case and other nonidentity cases just as nicely as the concept of expected value does—which is to say, quite nicely.

10. A Problem for Probable Value: The All-But-Known Success and the Unexpected Really Bad Outcome

At the same time, issues remain. Consider a new case, All-But-Known *Success*, a classic case that seems to cry out for the concept of expected value. Obviously, however, adding still a fourth necessary condition rooted in expected value won't work since we would then get false results in All-But-Known Disaster. So what are we to do with cases like All-But-Known *Success*? In this new case, the probabilities are clearly relevant. But by what principle are they to be brought to bear?

Graph 4. All-But-Known Success

		0.9999	0.0001
c1 undertakes mission	+10 +1 +0 -1000	*o1* **p** PV: 9.999	*o2* **p** PV: -0.1
c2 doesn't undertake mission	+10 +1 +0 -1000	*o3* **p** PV: 0.9999	*o4* **p** PV: 0.0001

Again, to add color to the case, let's suppose still another mother-and-child scenario—and another one-off case. Here, the woman must choose whether to undertake a mission to rescue her only child p from space aliens who have confined p in a cage allowing p only the minimal conveniences of life (nutrition, hydration and occasional ablution) or to do nothing and leave p to live out a life only *barely* worth living. If the woman's mission succeeds, p will be restored to a wellbeing level of approximately what the woman herself will have and what others in her and her child's generational cohorts will have. Because of the woman's overall proficiency and vast experience in previous extremely dangerous search-and-rescue efforts, the chances that her mission will succeed is extremely high. On the other hand, if her mission fails, the space aliens are sure to become aware of her efforts (her transport vehicle leaving a signature trace in the alien atmosphere, easily recorded by their sophisticated and highly sensitive monitoring devices) and they will (certainly) respond by torturing p in horrible ways for the remainder of p's very long natural life, leaving p with an existence far less than one worth having.

Let's suppose, too, that the woman chooses to proceed with the mission—chooses, that is, c1. And let's suppose as well that, against all odds, the mission fails, and p is left to suffer in the horrible ways we have just described.

This is the sort of scenario that reminds us that to *all but know* something is not to *know* something.

The question is: both AV and PV being relatively low, and the existence condition being satisfied, and EV being off the table, on what grounds (if any) do we say that

the woman's choice of c1 is permissible *even though* it has ended in disaster for p?

As already underlined, the consolidated principle PVMP+AVMP+EC, offering mere *necessary* conditions on wrongdoing, has no capacity, for any choice, to generate the result that that choice is wrong. It can at most generate, for a given choice, that that choice is permissible.

That means that the problem for the combination principle isn't that it generates a *false* result—the result that c1 is wrong—but rather that it's *incomplete*: that it *doesn't* generate the result that c1 is permissible.

It's—again—the probabilities that are at stake that make us want to say that c1 is permissible—that in All-But-Known Success, the mother all-but-knew, prior to choice, that things would work out well for p. But by what principle are the probabilities brought to bear to determine that c1 at o2, where disaster has struck, is also permissible? We can't appeal to the fact that PV or AV is maximized or to p's nonexistence, and EV is off the table. But what else is there?

11. Two options: Appeal to epistemic security or accept choice as wrong

To address this issue, we seem to have two options. The first exploits the fact that the most distinctive feature of All-But-Known Success is the *very* high probability of success (.9999) in that case in combination with the fact that p has a lot at stake (will p have a life well worth living, or a life just barely worth living?).

But what's the principle?

It seems that we can, in advance of a principle that we can actually test, say some things that seem both relevant and correct. We can say that, when the agent's position is *epistemically secure* relative to the better outcome and the agent all-but-knows that things will turn out well, as in All-But-Known Success, then the permissibility that comes with a choice that ends in an outcome that *does* turn out well can be *attributed* to the choice that ends in an outcome that (against all odds) *doesn't* turn out well. In contrast, when the agent's position is *epistemically weak* relative to the successful outcome, any such attribution of permissibility would be inappropriate. For in that sort of case, the agent can't plausibly be said to all-but-know, much less *know*, that things will turn out well. Ditto All-But-Known Disaster.

This isn't, of course, a principle we can test; it's just a bare-boned description of a schematic for such a test, one that links moral evaluation not just to the usual suspects, that is, facts about outcomes and the probabilities that those outcomes will obtain under the choices under scrutiny, but also to the agent's particular epistemic state. From there, the question remains whether an appropriate fourth necessary condition on wrongdoing can be fashioned.

The second is to take the position that we aren't challenged, in this case, to find grounds for the claim that the choice that ends in disaster—here, c2 at o2—is permissible. We should be comfortable with the fact that the door is open to a finding of wrongdoing in this particular case. For perhaps in this particular case the choice (though it *looked* permissible starting out) really *was* wrong.

References

Broome, John (1992). *Counting the Costs of Global Warming*. Cambridge: White Horse.

Kavka, Gregory (1981). "The Paradox of Future Individuals." *Philosophy & Public Affairs* 11: 93–112.

Parfit, Derek (1987). *Reasons and Persons*. Oxford University Press (originally published 1984).

Roberts, M.A. (2007). "The Nonidentity Fallacy: Harm, Probability and Another Look at Parfit's Depletion Example." *Utilitas* 19: 267–311.

Roberts, M.A. (2009). "The Nonidentity Problem and the Two Envelope Problem." In *Harming Future Persons*, eds. M.A. Roberts and D. Wasserman. Springer, pp. 201–228.

Roberts, M.A. and David Wasserman (2016) "Dividing and Conquering the Nonidentity Problem." In *Current Controversies in Bioethics*, eds. Matthew Liao and Collin O'Neil. Routledge, pp. 81–98.

Wlodek Rabinowicz[1]

Getting Personal—The Intuition of Neutrality Re-interpreted

According to the Intuition of Neutrality, there is a range of wellbeing levels such that adding people with lives at these levels doesn't make the world either better or worse. As lives in the neutral range can be good for those who live them, this intuition is in conflict with one of the main tenets of welfarism; it creates a disparity between what is good for a person and what is impersonally good. Adding a person with a good life needn't make the world better. In "Broome and the Intuition of Neutrality" (2009) I suggested, but did not elaborate, a re-interpretation of the neutral range that would remove the problematic disparity. On this re-interpretation, a life at a level within the neutral range is not merely impersonally neutral; it is also neutral in its personal value: neither better nor worse for its owner than non-existence. Nevertheless, among such personally neutral lives, some might still be personally better or worse than others, provided that they are incommensurable in their personal value with non-existence. In this paper, I explore some of the implications of this 'personalization' of the Intuition of neutrality. In particular, I discuss its worrisome implications for neutral-range utilitarianism (NRU). While NRU was originally proposed as a way to avoid the Repugnant Conclusion, it turns out this conclusion is re-instated on the new interpretation and, contrary to what was suggested in my 2009-paper, it remains repugnant. A related point is that it no longer holds that all personally good lives must be better for a person than personally neutral lives. Nor that all personally bad lives must

[1]Department of Philosophy, Lund University, wlodek.rabinowicz@fil.lu.se.

be worse than personally neutral lives. While this might seem strange, it should be accepted. As for the worrisome implications of NRU, these implications do not undermine the personalized Neutrality Intuition itself. The latter might well be retained even if NRU is given up.

*

According to the Intuition of Neutrality, in its axiological version, there is a range of wellbeing levels such that adding people with lives at these levels doesn't make the world either better or worse. On the standard interpretation of this 'neutral range', it extends from the zero level of wellbeing upwards, with the upper limit being some positive, though not very high, wellbeing level. (On the radical interpretation, the range of neutrality still starts at zero but has no upper limit.) Since a life at a positive wellbeing level is thought to be good for the person who lives this life, the Intuition of Neutrality drives a wedge between what is good for a person and what is impersonally good: Adding a person with a life that is good for her but has a wellbeing level within the neutral range doesn't make the world better, even if no one else is negatively affected by the addition. To this extent then, the Intuition comes into conflict with one of the basic tenets of welfarism.

In Rabinowicz (2009) I discussed the Intuition of Neutrality and defended it against John Broome's challenging objections (Broome 2004). I also sketched a particular axiological theory – neutral-range utilitarianism – that incorporates this intuition. While Broome's criticisms did not target the disparity the Intuition brings in between personal and impersonal goodness,[2] I still thought that it was there the main problem with the Intuition was to be found. Therefore, I also suggested, but did not elaborate in that paper, a re-interpretation of the neutral range that would remove the problematic disparity. On this re-interpretation, a life at a level within the neutral range is not merely impersonally neutral – it does not merely fail to make the world better or worse – but it also is neutral in its personal value: It is neither good nor bad for a person to have such a life. Its wellbeing level is neither positive nor negative. A life at this level is thus neither worth living nor worth not living. To

[2] Indeed, Broome's own favourite theory in that book, critical-level utilitarianism, also implies such a disparity. In fact, it goes even further in this respect than the Intuition of Neutrality: it implies that what is personally good might be impersonally bad. Adding a person with a life that is good for her makes the world worse if the wellbeing level of the added life, while positive, is lower than the critical level adopted by the theory. To be sure, Broome also suggests that the location of the critical level might well be indeterminate. This would make it indeterminate, for some positive wellbeing levels, whether adding people with lives at those levels makes the world better or worse. Indeterminacy makes the disparity between what's good for a person and what's good for the world less blunt. But it doesn't remove it.

put it differently, it is neither better nor worse for its owner than non-existence. Nevertheless, among such personally neutral lives, some might still be personally better or worse than others. Thus, there might be a range of personally neutral well-being levels – some higher and some lower. This is possible provided that lives at these levels can be incommensurable in their personal value with non-existence. Unlike equal goodness, incommensurability is not a transitive relation, which means that a personally neutral life – a life that is incommensurable with non-existence – can still be personally better than another personally neutral life.

In this paper, I want to explore some of the implications of this 'personalization' of the Intuition of neutrality. In particular, while such a move might seem to make neutral-range utilitarianism more plausible, I will argue that the appearances are misleading. For one thing, the resulting theory gets considerably more complicated: We not only need to allow for lives that are incommensurable with non-existence; we also need to allow for lives that are incommensurable with each other. This means that some lives' wellbeing levels might not be ordinally comparable. Allowing for lives that are incommensurable in personal value makes technical trouble for a utilitarian axiology. It is not even obvious that such an axiology can accommodate incommensurable lives. While this worry, I think, can be dealt with, there is also a substantive issue that neutral-range utilitarianism needs to confront. In my 2009-paper, I pointed out that personalizing the Intuition of Neutrality re-instates the Repugnant Conclusion that neutral-range utilitarianism was supposed to avoid. But I also suggested that this re-interpretation at the same time removes, or at least assuages, the repugnancy of the Repugnant Conclusion. I now think this diagnosis was premature: If the Intuition of Neutrality is re-interpreted on the lines I have suggested, then even a genuinely repugnant conclusion can be re-instated. This poses a challenge to neutral-range utilitarianism. Indeed, the whole landscape of personal value becomes more complicated on this new picture. As suggested above, we need to give up the standard assumption that the wellbeing levels of different lives are linearly ordered. As I am going to show, one of the implications of this change is that it no longer holds that all personally good lives must be better than personally neutral lives. Nor that all personally bad lives must be worse than personally neutral lives. But while such implications of this new approach are surprising, they should be accepted. As for the issues that on this new interpretation arise for the neutral-range utilitarianism, they do not undermine the personalized Neutrality Intuition itself. The latter can be kept even if the neutral-range utilitarianism were given up.

In section 1, I will present the standard interpretation of the Intuition of Neutrality and then move on, in Section 2, to neutral-range utilitarianism (NRU). In section 3, I will describe the personalized version of the Intuition and the way this

personalization affects NRU. The picture I will draw will be essentially the same, although more elaborate, as the one I have sketched in Rabinowicz (2009). Then, in section 4, I will consider how this picture changes and gets more complicated, in ways I didn't envisage in my earlier paper, if one allows that different lives might not be commensurable in their personal value.

One important part of morality is concerned with what is better or worse for people. According to a popular version of this *person-affecting* idea of morality, what is better (worse) must be *better (worse) for* someone.[3] However, it is unclear how we are to put this idea to use in non-identity cases, i.e., cases where, depending on what we decide to do, different people will come to exist in the future. Indeed, there seems to be a clear tension between the person-affecting idea and some of our considered judgements about non-identity cases. In at least some non-identity cases we want to say that one outcome is better (or worse) than another in virtue of the wellbeing of people who do not exist in both. For example, we want to say that creating a very unhappy person makes the world worse, other things being equal. But how can we say this, if an outcome is worse only if it worse for someone? In order to comply with a person-affecting morality in this case, we need to show that coming into existence can be worse for a person. But can it really be worse for a person to exist than not to exist, and thus better for her not to exist than to exist? That seems to require that the person would have been better off not existing, which sounds paradoxical.

In this paper, I am going to discuss some recent attempts to ease this tension. According to these attempts, we can stick to a person-affecting morality and still avoid the counterintuitive judgement that no outcome is better or worse in virtue of the wellbeing of people whose existence is contingent on our choice. I shall show that none of these attempts is convincing. That leaves us with only one option: to reject the person-affecting constraint in its current form.

In section 2, I shall say more about non-identity cases, and list the most morally salient ones. In section 3, I shall make more precise what a person-affecting morality amounts to. In section 4, I shall present an argument that spells out the tension between person-affecting morality and our judgements about non-identity cases. The argument's conclusion is that no outcome can be better or worse than another in terms of the well-being of people who do not exist in both. In sections 5 to 10, I shall discuss possible ways to resist this argument while sticking to a person-affecting morality. I shall especially focus on the approach recently defended by the so-called 'Scandinavian existentialists'.[4] I shall argue that the main problem with

[3] See Temkin (1993a), (1993b), and Holtug (1996). The label "Person-Affecting Restriction" was introduced by Glover (1977), p. 66, but see also Narveson (1967).

[4] See, for instance, Arrhenius & Rabinowicz (2010), (2015), Johansson (2010), and Holtug (2001). See also, Adler (2009), and Adler (2011) for similar ideas. Some seeds for this approach seem to have been planted already in Parfit (1995), appendix G, p. 490.

their approach is that they fail to fully acknowledge what it means to say that an abstract state of affairs has value.

1. The Intuition of Neutrality—standard interpretation

As a normative principle of action, the Intuition of Neutrality goes back to Jan Narveson's famous pronouncement: "We are in favour of making people happy, we are neutral about making happy people." (Narveson 1973) A more guarded, less committal formulation of this ethical Intuition is provided by John Brome in *Weighing Lives*:

> We think intuitively that adding a person to the world is very often ethically neutral. We do not think that just a single level of wellbeing is neutral [...].
>
> (Broome 2004, p. 143)

Here, following Broome (ibid. p. 145f), I am going to focus on the axiological version of the Intuition. It can be put as follows:

> *Intuition of Neutrality*: There is a range of wellbeing levels, call it *the neutral range*, such that adding a person with a life at one of these levels, without affecting the wellbeing of anyone else, does not make the world either better or worse.

Broome himself was initially attracted to this Intuition, but then eventually felt compelled to give it up, for several reasons that I am not going to discuss in this paper. But see Rabinowicz (2009) for an extended critical discussion of Broome's objections. For Broome's reply, see Broome (2009).

What is supposed to be the scope of the neutral range? On the moderate interpretation, this range begins at the zero level of wellbeing and extends upwards, to some positive, though not too high, wellbeing level. On the radical interpretation, it extends all the way up, to infinity.[5] Here, I will focus on the moderate interpretation, according to which the neutral range has an upper bound as well as a lower bound. On this view, adding bad lives – lives at negative wellbeing levels – makes the world worse, while adding excellent lives makes the world better. In what follows, I am

[5] Cf. Broome (2004, p. 144): "Some people think this range is infinitely wide. They think that a person's existence is neutral, however good her life would be if she did exist. It is not neutral if her life would be bad, so there is a lower boundary to the neutral range. But there is no upper boundary. That is one view. A more moderate view is that the range has both an upper and a lower boundary, but there is nevertheless a range of neutral lives in between."

going to refer to this interpretation of the neutral range as the 'standard' one. Though I don't want to imply by this label that any dominating view has already developed in this area.

By a bad life I mean a *personally* bad life, i.e. a life that is bad for the person who lives it. For that person such a life is 'worth not living': It is worse for her than non-existence. I assume that lives at negative levels of wellbeing are bad in this sense.

Analogously, a good life is a life that is personally good, or worth living: it is a life that is better for the person who lives it than non-existence. A life's wellbeing level is positive iff the life in question is good in this sense.

In line with this interpretation of positive and negative wellbeing levels, the zero level of wellbeing characterizes a life that for the person who lives it has the same value as non-existence.

I will assume that a life's personal value does not depend on the numerical identity of the person who lives this life. Thus, for anyone who would live this life its personal value would for her be the same. I think it is a reasonable assumption, provided we take a person's life to contain everything that characterizes this person – not merely what she does and what happens to her, and not just her external circumstances as they change over time, but also her internal, psychological make-up and internal history. Given this all-encompassing conception of a life, it is plausible to assume that, for anyone who would live it, it would have the same value.

Several philosophers have argued that personal value comparisons between one's life and one's non-existence make no sense. It makes no sense on their view to suggest that it can be better (or worse) for me to live my life than never to live at all. I disagree; I think that comparisons of this kind do make sense, even though they might be difficult to make. (See Arrhenius and Rabinowicz 2010, 2015; cf. also Johansson 2010 and Holtug 2001 for related views. For challenging objections, see especially Bykvist 2007, 2014.). Here I cannot argue this point, but I can at least say something to allay the immediate worry such a claim might invite.

Consider:

> (i) It is better for John to have the life he has than not to exist.

This statement implies, according to the critic, that

> (ii) If John didn't exist, it would have been worse for him than to have the life he has.

But, if John didn't exist, nothing *could* have been worse, or better, for him, as there would be no him for whom anything could be better or worse. Non-existents cannot have any properties or stand in any relations. Thus, (ii) is absurd, which shows that such claims as (i) have absurd implications.[6]

In my view, this objection is not justified: Contrary to appearances, (ii) does not follow from (i). (i) and the consequent of (ii) state that a certain relation obtains / would have obtained between three relata: John and two states of affairs, John having the life he has and John's non-existence. Now, a relation can only obtain if all its relata exist. Consequently, it couldn't have obtained if John did not exist. Which means that (ii) must be false: If John did not exist, no relation in which he is one of the relata could have obtained. On the other hand, (i) may well be true. If John does exist, the requisite relation between him and the two states of affairs may well obtain.

But, one might wonder, if John does exist, then what about the state of his non-existence? Does this state exist in such a case? The answer is yes, if we think of states of affairs as abstract objects. As such, they exist even when they do not obtain.[7] We may therefore conclude that there is no implication from (i) to (ii): (ii) is necessarily false for the reasons that do not apply to (i).

I will therefore continue to account for the personal value of a life in terms of comparisons with non-existence: For any person, a life L is good (bad, neutral) for that person iff it would be better (worse, neither better nor worse) for her to have life L than not to exist at all.

Let us go back to the main thread. The Intuition of Neutrality together with the standard conception of the neutral range imply that adding a good life need not always make the world better. It will not make it better, or worse for that matter, if the wellbeing level of that good life lies within the neutral range. Thus, on this view, there is a striking *disparity* between what is good for a person and what is impersonally good: A life might be good for the person who lives it without being impersonally contributively good – without making the world better. This disparity might well be precisely what attracts some philosophers to the Intuition of Neutrali-

[6] Cf. Broome (1999, ch. 10, p. 168): "[I]t cannot ever be true that it is better for a person that she lives than that he should never have lived at all. If it were better for a person that she lives than that she should never have lived at all, then if she had never lived at all, that would have been worse for her than if she had lived. But if she had never lived at all, there would have been no her for it to be worse for, so it could not have been worse for her." See also Parfit 1991 [1984], p. 395. In *Weighing Lives*, though, Broome is less categorical: He recognizes that comparing a person's life with her non-existence, in terms of its value for the person in question, might possibly be made sense of after all (see Broome 2004, p. 63).

[7] Note that only if states of affairs can exist without obtaining can there be any relations between states that do not co-obtain. Incompatible states cannot co-obtain. Thus, not even the relation of incompatibility would obtain between them if they couldn't exist unless they obtained. For if one of them obtains, the other does not.

ty, but at the same time it is a worrying implication from a strictly welfarist point of view. I will return to this issue in Section 3.

How is the neutrality of life additions to be interpreted? If adding a life with a wellbeing level in the neutral range (while keeping the wellbeing levels of everyone else constant) doesn't make the world either better or worse, will the world with such an added life be equally as good as the original world? Or will it instead be *incommensurable* with the original world – neither better or worse nor equally as good? It is easy to prove that incommensurability is the only viable alternative.

As a preparation for the proof let us note that the neutral range is supposed to contain more than just one wellbeing level. Furthermore, according to the standard conception of the neutral range. the wellbeing levels are linearly ordered, from higher to lower. Consequently, it follows that *for very wellbeing level m in the neutral range there is at least one level in that range that is higher or lower than m*. (While the linear ordering assumption will be criticized in section 4, the italicized claim won't be questioned. And it is only this weak claim that is needed for the proof to follow.)

Let *A* be the original world and *B* the world with an added person, call her Barbara, whose life in *B* is at a wellbeing level *m*. Suppose that *m* lies in the neutral range. By the Intuition of Neutrality, this implies that *B* is neither better nor worse than *A*. We want to prove that *B* is not equally as good as A either. It will then follow that these two worlds must be incommensurable.

Proof:[8] As shown above, there must be at least one level *n* in the neutral range that is either higher than *m* or lower than *m*. Now, consider a world *C* in which Barbara is added to the original world at wellbeing level *n*. Just as in *B*, no one else in *C* is affected by this addition.

Case 1: $n > m$. Since it is better for Barbara to live at a higher rather than a lower level of wellbeing, *C* is better for her than *B*. And it is equally as good as *B* for everyone else. We may therefore conclude that *C* is better than *B*. This is implied by the following general principle:

Suppose that worlds *X* and *Y* have the same population, *I*. (i) If *X* is better than *Y* for some individuals in *I*, while it is equally as good as *Y* for everyone else in *I*, then *X* is better than *Y*. (ii) If *X* and *Y* are equally as good for everyone in *I*, then *X* and *Y* are equally good.

This Pareto-like principle is an important part of the welfarist outlook. It establishes a minimal connection between personal and impersonal good. Following Broome (2004, section 8.2) we might call it the *Principle of Personal Good*.

This principle is compatible with the Intuition of Neutrality, even though the latter, on the standard interpretation of the neutral range, introduces a disparity

[8] In its essentials, this proof is due to Broome; cf. his (2004), pp. 146ff.

between personal and impersonal good. While the Intuition of Neutrality focuses on value comparisons between worlds with partly different populations, the Principle of Personal Good – as stated above – is restricted to comparisons between worlds that have the same population. This restriction is crucial. Without it, the Principle of Personal Good would come into conflict with the Intuition of Neutrality. To see this, note that if $n > m$ and both m and n are in the neutral range that stretches upwards from zero, n must be a positive level of wellbeing. But then *C*, in which Barbara's level is n, is better for her than *A*: It is better for her to live a life at a positive wellbeing level than not to exist at all. At the same time *C* is equally as good as *A* for everyone else. Thus, in the absence of the restriction, the Principle of Personal Good would imply that *C* is better than *A*. However, since n lies in the neutral range, the Intuition of Neutrality implies that *C* is not better than *A*.

It is an interesting issue whether and how principles such as the Principle of Personal Good can be extended to comparisons between worlds with variable populations. The argument I have just given shows that the straightforward extension of this principle would be in conflict with the Intuition of Neutrality *if* the standard conception of the neutral range is assumed. This argument doesn't go through, however, if the neutral range is re-interpreted as will be done below, in section 3. Indeed, the straightforward extension of the Principle of Personal Good will then become possible without getting into conflict with the re-interpreted Intuition of Neutrality: The restriction to worlds with the same population will not be needed.

But let us return to the main proof. Suppose, for *reductio*, that *B* is equally as good as *A*. Then *C*, which by the Principle of Personal Good is better than *B*, must be better than *A*. (Betterness is transitive across equal goodness.) But since n just as m lies in the neutral range, the Intuition of Neutrality implies that *C* is *not* better than *A*. We must therefore conclude that *B* and *A* are not equally good.

Case 2: $n < m$. The Principle of Personal Good now implies that *B* is better than *C*. Consequently, if *B* were equally as good as *A*, *A* would also be better than *C*. But this again is excluded by the Intuition of Neutrality.

This completes the proof.

Thus, adding a person with a wellbeing level that lies in the neutral range creates an incommensurability. But how is this incommensurability to be explained? In the next section I suggest an explanation.

2. Analysis of value relations and neutral-range utilitarianism

I will start this section with rehearsing my general proposal as to how one might analyse different value relations, including the relation of incommensurability. (Cf.

Rabinowicz 2008, 2012.) I will then suggest, following Rabinowicz (2009), how this account can be applied to the problem at hand.

My account of value relations adheres to the general format of the fitting-attitudes analysis of value (FA-analysis). On this approach, value statements are interpreted as normative assessments of pro- and con-attitudes towards evaluated objects. As for statements of value relations, it is then natural to interpret them as normative assessments of preferences regarding the compared objects. Betterness and equal goodness are analyzed as follows:

> x is *better* than $y =_{df}$ x ought to be preferred to y.

> x is *equally as good* as $y =_{df}$ x and y ought to be equi-preferred.

Consequently, items x and y are incommensurable iff none of them of them ought to be preferred to the other nor ought they be equi-preferred.

There are two levels of normativity, the strong level of requirement ('ought') and the weak level of permission ('may'). Allowing for weak normativity as regards preferences makes room for further types of value relations. In particular, we can define the notion of parity, which is the typical (though not the only one) form of incommensurability:

> x and y are *on a par* $=_{df}$ x may be preferred to y and y may be preferred to x.

Thus, in cases of parity, opposing preferences regarding the compared items are permissible.

This approach to value relations can easily be formalized. In Rabinowicz (2008), I proposed the following *intersection modelling*:

The modelling has two components: the domain D of items that are compared and the class K of all permissible preference orderings of this domain.

Betterness is defined as required preference:

> x is *better* than y iff x is preferred to y in every ordering in K.

This is just another way of saying that x ought to be preferred to y: every permissible preference ordering of the domain must incorporate this preference.

Analogously, equal goodness is defined as required equi-preference:

> *x* and *y* are *equally good* iff they are equi-preferred in every *K*-ordering.

In case of parity, opposing preferences are allowed:

> *x* and *y* are *on a par* iff *x* is preferred to *y* in some *K*-orderings and *y* is preferred to *x* in some other *K*-orderings.

Betterness is a transitive and asymmetric relation and equal goodness is transitive and symmetric. In addition, betterness is transitive across equal goodness. To guarantee that these conditions hold, we need to impose certain minimal formal constraints on the preference orderings in *K*. *x* will be said to be *weakly preferred* to *y* iff *x* is either preferred to *y* or equi-preferred with *y*. Indeed, we might just as well take this relation of weak preference to be our primitive concept and then define preference and equi-preference as, respectively, the asymmetric and the symmetric parts of weak preference: *x* is preferred to *y* iff *x* is weakly preferred to *y* but not vice versa; *x* is equi-preferred with *y* iff *x* is weakly preferred to *y* and vice versa. What we then need to assume about the orderings in *K* is that in each of them weak preference is reflexive and transitive. This constraint on permissible orderings is what gives us the desired formal properties of betterness and equal goodness.

Now, how can we apply this general modelling to the problem at hand? As the domain *D* we now take the set of possible worlds. If we accept the basic tenet of welfarism – the view that the value of a world is fully determined by the wellbeing levels of the individuals in that world – we can, for simplicity, identify each world with a wellbeing distribution: an assignment of (lifetime) wellbeing levels to the individuals that exist in this world. The question now is what the class *K* of permissible orderings of this domain is supposed to look like. Specifying *K* will determine the value relations that obtain between the possible worlds in *D*.

To specify *K* is thus to provide a substantive axiology: more precisely, a substantive population axiology. As is well-known, population ethics is haunted by conflicting intuitions and impossibility results. (Cf. Arrhenius 2000, 2011, 2016, and forthcoming. For a critical discussion, cf. Carlson, this volume.) It is a fair conjecture that no axiology can accommodate all plausible principles concerning value comparisons between worlds with variable populations. Here, I will focus on a utilitarian axiology, not because I find it fully satisfactory, but because it is a relatively plausible and at the same time a very simple form of welfarism. It is a simple and definite theory to work with. However, the kind of utilitarian axiology I want to consider has to make room for incommensurabilities in world comparisons, in order to accom-

modate the Intuition of Neutrality. In Rabinowicz (2009) I called it *neutral-range utilitarianism*.

Let *w*, *w'*, ... stand for lifetime wellbeing levels. We assume, until further notice, that these levels are measurable on a ratio scale: Thus, only the unit of measurement is arbitrarily chosen. (This strong measurability assumption will be relaxed in Section 3 and then much further relaxed in Section 4.)

Now, one way to think of the neutral range is that wellbeing levels in that range are potential candidates for being permissible 'critical levels' – permissible benchmarks – for preference. It is permissible to choose any of them, say level *w*, as the maximum below which adding new lives to the population is dispreferred. If a wellbeing level *w* lies in the neutral range, it can thus be used as a benchmark for determining the position of each world in the preference ordering. This position is obtained by summing up, for all individuals in the world in question, the surpluses and the deficits in their wellbeing, as compared with the chosen benchmark. Thus, let $I(A)$ be the set of individuals that exist in a world A and let w_{iA} be the wellbeing level of an individual i in A. The sum of wellbeing surpluses and shortfalls in A, relative to benchmark w, equals

$$\sum_{i \in I(A)} (w_{iA} - w).$$

The higher this sum is, the higher the world ends up in the preference ordering determined by *w*. Any two worlds for which this sum is the same occupy the same position in the ordering.

In this way, different preference orderings in K correspond to different choices of critical levels from the neutral range. A choice of a particular level specifies the point beyond which it is preferred to add a life to the population and below which such addition is dispreferred. Different choices of critical levels *w* from the neutral range (though only from that range) are all admissible and each such *w* induces a permissible preference ordering P_w on the set of worlds. Note that every P_w is a complete ordering: it contains no gaps. That is, for any two worlds, one of them is preferred in P_w to the other or they are equi-preferred in P_w. Class K might also include gappy preference orderings; it should be permissible to have incomplete preferences between worlds. I will return to this possibility in a moment. But if we disregard it, we can define the relations of betterness, equal goodness and parity between worlds as follows:

> A world A is *better* than a world B iff for all *w* in the critical range, A is preferred to B in P_w.

> *A* is *equally as good* as *B* iff for all *w* in the neutral range, *A* is equi-preferred with *B* in P_w.
>
> *A* is *on a par* with *B* iff for some *w* and *v* in the neutral range, *A* is preferred to *B* in P_w and *B* is preferred to *A* in P_v.[9]

To illustrate, consider again the world *B* in which Barbara is added to the original world *A* at level *m*, where *m* belongs to the neutral range. Since this range contains more than one wellbeing level and since levels are linearly ordered from higher to lower, there must be some level *n* in that range that is higher or lower than *m*. Suppose that $n > m$. Then *A* is preferred to *B* in P_n but not in P_m. (In P_m, *A* and *B* are equi-preferred.) Analogously, if $n < m$, then *B* is preferred to *A* in P_n but not in P_m. Thus, in either case, *K* will contain preference orderings that differ from each other in how they rank *A* and *B*. This means that *B* is incommensurable with *A*, neither better nor worse than the latter, nor equally as good. Indeed, if the neutral range contains both levels higher than *m* and lower than *m*, *A* and *B* are on a par.

What about gappy preference orderings in *K*? Let *W* be any non-empty subset of wellbeing levels in the neutral range. For example, *W* might be a sub-interval of that range. Let P_W be the set of complete preference orderings induced by the wellbeing levels in *W*. Plausibly, the intersection of P_W – the common part of the orderings in P_W – is also a permissible preference ordering of worlds.[10] This intersection $\cap P_W$ will contain gaps if *W* contains more than one wellbeing level. Intuitively, $\cap P_W$ represents preferences of someone who is undecided between levels in *W* as potential benchmarks – someone who has not made up his mind as to where exactly to draw the line between preferred and dispreferred life additions.

Introducing gappy orderings into *K* in this way does not affect the extensions of the four typical value relations between worlds: better, worse, equally as good, and on a par. Nor does it affect the incommensurability relationships between worlds. Therefore, in most cases, there's no need to consider these incomplete *K*-orderings.

The population axiology described above is what I call *neutral-range utilitarianism* (NRU, for short). It combines total-sum utilitarianism with the Intuition of Neutrality. This axiology is formally identical with the theory that has been put forward by Blackorby, Bossert and Donaldson (1996). They call it the "incomplete critical-level utilitarianism" (ICLU). ICLU is a generalization of the more familiar

[9] In typical cases of parity, there will also be some *u* such that A is equi-preferred with B in P_u. Indeed, this will hold in all mere-addition cases I here discuss. But in the interest of greater generality I abstain from imposing this condition as part of the definition of parity.

[10] In this intersection of P_W, world *A* is weakly preferred to world *B* iff *A* is weakly preferred to *B* in every ordering in P_W.

critical-level utilitarianism (CLU), originally proposed by Blackorby and Donaldson (1984). CLU is a utilitarian theory that picks out a specific non-negative wellbeing level w as the unique critical level and then lets the value ordering of worlds coincide with P_w. Thus, on CLU, world A is better than world B iff $\sum_{i \in I(A)}(w_{iA} - w) > \sum_{i \in I(B)}(w_{iB} - w)$. If these two sums are equal, the two worlds are equally good. CLU entails that the value ordering of worlds is complete: there are no incomensurabilities. As this completeness of evaluation might well be questioned, ICLU has been proposed as a less categorical option. On ICLU, the value ordering of worlds is taken to be a compromise between different complete value orderings that are generated by different choices of critical levels. As such a compromise, it retains what is common to the alternative complete evaluations and leaves gaps at places where the complete evaluations disagree.

Clearly, even though ICLU and NRU are structurally identical theories, they differ in their philosophical motivations. On ICLU, the ultimate value ordering is a compromise between different complete value orderings, while on NRU it is instead the intersection of permissible *preference* orderings. NRU is based on the analysis of value relations in terms of permissible preferences. This anchoring in the FA-format of analysis is absent in ICLU. Indeed, the philosophical motivation for ICLU has never been made very clear by its proponents and it is therefore probably not accidental that in their subsequent publications Blackorby, Bossert and Donaldson have reverted to CLU as their favoured axiology.

NRU is also structurally identical to Broome's (2004) version of CLU. On this version, while there is just one critical level, its precise location is *indeterminate*. Instead of the neutral range we thus have a zone of indeterminacy – the set of wellbeing levels such that it is indeterminate which of them is the critical level but determinate that it is one of them. From this zone of indeterminacy, different precisifications of the theory choose different levels as the critical one and offer complete evaluations of possible worlds based on these choices. In the standard supervaluationist manner, what is common to all the precisifications (i.e., their intersection) is determinately true according to this axiology. Statements that hold on some precisifications but not on others are neither true nor false. They have an indeterminate truth value. Statements that don't hold on any precisification are false. On Broome's proposal, there are thus no incommensurabilities between worlds; instead, we have indeterminacies concerning their mutual value relations. Again, this axiology, in spite of its close structural similarity to NRU, has a different philosophical interpretation.

Let's move on. We have seen how NRU is supposed to work. But how plausible is it as a population axiology?

As is well known, the standard (total-sum) utilitarianism implies the Repugnant Conclusion (see Parfit 1991[1984], ch. 17):

> *Repugnant Conclusion*: For any world whose inhabitants have excellent lives, there is a better possible world all whose inhabitants have lives barely worth living – lives that are good but only barely so.

Call the former and the latter world the Happy World and the Drab World, respectively. If the Drab World has a sufficiently large population, its total sum of wellbeing will exceed that of the Happy World. Increases in population size compensate and indeed outweigh losses in life quality.

Unlike total-sum utilitarianism, NRU does not allow for such facile compensations. Barely good lives, i.e. lives at very low positive levels of wellbeing, have wellbeing levels within the neutral range. On the standard conception of that range, it stretches from zero upwards to some relatively high (though not too high) level of wellbeing. Thus, adding lives with low positive levels of wellbeing to the world does not make it better. On NRU, the Repugnant Conclusion is avoided.

What is not avoided is the Weak Repugnant Conclusion – a claim that is just like the Repugnant Conclusion, but with "better" replaced by "not worse". If the Drab World contains sufficiently many people, then, on NRU, it will be incommensurable with the Happy World. Indeed, the two worlds will be on a par. On some choices of a benchmark w from the neutral range, very close to zero (closer than the level of drab lives), the Drab World will be preferred to the Happy World, while choosing a higher w will yield the opposite preference. While Weak Repugnant Conclusion might be hard to accept, it is not as outrageous as the original Repugnant Conclusion.

Apart from the Repugnant Conclusion, NRU also avoids the so-called Sadistic Conclusion, which plagues CLU, the critical-level utilitarianism (see Arrhenius 2000):

> *Sadistic Conclusion*: For any world whose inhabitants have terrible lives, there is a worse possible world whose inhabitants all have lives worth living (if only barely so).

CLU has this sadistic implication because the wellbeing levels of some lives worth living are lower than the critical level. Thus, on CLU, each such life detracts from the value of the world. Consequently, if the population of the Drab World is sufficiently large, this world will be worse than the Terrible World.

NRU avoids the Sadistic Conclusion if the neutral range goes all the way down to the zero level of wellbeing, as the standard interpretation of this range would have

it. Low positive wellbeing levels lie within the neutral range. Therefore, adding lives at such levels does not detract from the value of the world: it does not make the world worse.

But again, NRU does not avoid the Weak Sadistic Conclusion, i.e., the claim which is just like the Sadistic Conclusion, but with "worse" replaced by "not better". The Drab World will not be worse than the Terrible World, but if its population is large enough, it will not be better. It will be incommensurable (on a par) with the Terrible World. It is a worrying and highly implausible implication.

This in itself might be a sufficient reason to consider another, alternative population axiology. But there is a more basic reason as well: As it stands, the Intuition of Neutrality is problematic.

3. The Intuition for Neutrality re-interpreted

As we have seen, on the standard conception of the neutral range, the Intuition of Neutrality introduces a disparity, a hiatus, between personal and impersonal goodness. A life might be good for a person who lives that life, better than non-existence, but still the addition of such a life to the world might not be impersonally good: It might not make the world better. It doesn't make it better if the wellbeing level of the added life lies within the neutral range.

The disparity between what's good for a person and what's good for the world will be seen by some as an appealing feature of the Intuition of Neutrality, indeed, as the reason to adopt it in the first place. But others will consider it highly problematic; they will see it as foreign to the welfarist outlook.

I am now going to consider a re-interpreted version of the Intuition – one that removes the disparity. This new version of the Intuition is obtained by a reinterpretation of the neutral range: The idea is to identify it with the range of wellbeing levels at which a life is *neither good nor bad* for the person who lives or could live that life – neither better nor worse for her than non-existence. In other words, the *impersonal neutral range* – the range of wellbeing levels at which additions of lives make the world either better or worse – is now identified with the *personal neutral range* – the range of wellbeing levels at which a life is neutral for the person who lives (or could live) this life: levels at which this life is neither good nor bad for her. (Cf. Rabinowicz 2009, pp. 390f, for this suggestion. See also Gustafsson 2016, where this suggestion is adopted and elaborated.[11])

Note that this re-interpretation presupposes that such personal neutral range does exist, i.e., that there is more than one wellbeing level at which a life is personally

[11] Gustafsson uses a slightly different terminology, though, and, more importantly, he does not define the personal value of a life by comparing it to non-existence.

neutral. And, more specifically, it presupposes that personally neutral lives can be better or worse for us than other personally neutral lives, despite the fact that all such lives are neither better nor worse for us than non-existence. How is it possible? The answer should be obvious by now. It can only be possible if lives at these personally neutral levels are incommensurable in their personal value with non-existence.

The argument for this claim is analogous to the one I have presented in section 1 above, in connection with impersonal neutrality. Here, I give it in a simplified and condensed version. Thus, suppose that levels m and n are both personally neutral and $m < n$. By assumption, lives at these levels are neither better nor worse for the persons who live them than non-existence. Nor can any of these lives be equally as good for them as non-existence, as the following argument shows. One of these lives, the n-life, is personally better than the other, i.e., it is such that it would be better for a person to live this life than to live the other life. But this means that if the m-life were personally equally as good as non-existence, then the n-life would be personally better than non-existence, contrary to the assumption. (The argument here depends on personal betterness being transitive across personal equal goodness.) On the other hand, if the n-life were personally equally as good as non-existence, then the m-life would be personally worse than non-existence, again contrary to the assumption. Thus, both of these lives must be incommensurable in their personal value with non-existence. More generally, if every personally neutral life is better or worse than some other personally neutral life, as it must be if there is more than one wellbeing level in the personally neutral range and wellbeing levels are linearly ordered, then every such life must be incommensurable in its personal value with non-existence.

How is such incommensurability in personal value between a life and non-existence to be understood? To answer this question, I again need to appeal to the FA-format of analysis, but this time apply it to personal value. There is no clear consensus among FA-analysts as to how personal value – goodness for a person – should be understood. I will adopt the view that what is good for a person is what anyone who cares for her ought to wish or desire for her sake.[12] Admittedly, this suggestion is not very precise: both the notion of caring for someone and the notion

[12] A proposal roughly along these lines was put forward by Darwall (2002). For a competing FA-account, according to which what is good for a person is what *anyone* (and not just those who care for her) ought to favour for her sake, see Rønnow-Rasmussen (2011). (Cf also Rønnow-Rasmussen 2018.) The normative reach of the latter account is much wider than that of the former – too wide, I think. Cf. Taurek (1977, p. 304): "When I judge of two possible outcomes that the one would be worse (or better) for this person or this group, I do not, typically, thereby express a preference between these outcomes. Typically, I do not feel constrained to admit that I or anyone *should* prefer the one outcome to the other." Though even Taurek would agree, I suppose, that I should have this preference if I care for the person or group in question.

of wishing/desiring something for someone's sake would need clarification.[13] Hopefully, though, the main idea of this proposal is sufficiently clear. Extending it to personal value relations is straightforward:

> *x* is better for *i* than *y* iff anyone who cares for *i* ought to prefer *x* to *y* for *i*'s sake.
>
> *x* is equally as good for *i* as *y* iff anyone who cares for *i* ought to equi-prefer *x* and *y* for *i*'s sake.
>
> Incommensurability between *x* and *y* obtains for *i* whenever neither of these items is better for *i* than the other, nor are they equally as good for *i*.
>
> Parity for *i* between *x* and *y* – the most typical form of incommensurability in personal value – obtains when it is permissible for anyone who cares for *i* to prefer *x* to *y* for *i*'s sake and likewise permissible for them to have the opposite preference.

The idea is now that this kind of personal parity can obtain between a life and non-existence. For certain wellbeing levels – the ones in the personal neutral range – it is permissible, for the sake of a person who might live a life at this level, to prefer that she has this life rather than she does not exist and likewise permissible to have the opposite preference – permissible, that is, for anyone who cares for that person.[14]

Normally, in cases of parity, equi-preference with regard to the items on a par is also permissible. If it is permissible to prefer *x* to *y* and likewise permissible to have the opposite preference, then it should also be permissible to equi-prefer *x* and *y*. This also applies to cases in which a life is personally on a par with non-existence: it should be permissible, for the sake of a person who could live that life, not only to

[13] Just to give an example of an issue that might need clarification: Can one care for an individual *i* and desire something for *i*'s sake if *i* does not exist (but *could* exist)? In principle, it should be possible: *i* in these locutions appears in an intensional context; Neither caring for *i* nor desiring something for *i*'s sake is a relation in which *i* is one of the relata. (Just as, say, thinking of Pegasus is not a relation in which Pegasus is one of the relata.) But this would mean that something can be good for *i* or better for *i* even if *i* does not exist, contrary to what I have assumed above, in Section 1. Should we welcome this implication? It would make personal value comparisons between someone's life and her non-existence even easier than I have previously suggested. On the other hand, caring for someone or desiring something for her sake requires that one is at least able to identify that person, which is difficult in case of non-existents. So perhaps we should evade this problem and simply add to the analysis of what it means that something is good or better for *i* an extra condition that *i* does exist? This would be easy, but rather ad hoc.

[14] For a suggestion as to why opposing preferences can be permissible in cases like this, see the next section, footnote 15.

prefer one of these alternatives to the other, but also to equi-prefer them – to be indifferent between her having this life and her non-existence. This point will be of some importance below.

Let me regiment the terminology. I will say that a life is personally better than (worse than, equally as good as) another life iff it is better (worse, equally as good) for a person to have the former life than (as) the latter. Analogously, a life is personally better than (worse than, equally as good as) non-existence iff it is better (worse, equally as good) for a person to have this life than (as) not to exist. In the same vein, I will talk about personal incommensurability, or parity, between a life and non-existence, or between one life and another life.

We have three kinds of lives:

> A (personally) good life = a life that is personally better than non-existence.[15]
>
> A (personally) bad life = a life that is personally worse than non-existence.
>
> A (personally) neutral life = a life that is neither (personally) good nor bad.

The re-interpreted Intuition of Neutrality only applies to the (personally) neutral lives: According to the Intuition, adding such lives is impersonally neutral; it does not make the world either better or worse. This re-interpretation removes the disparity between personal and impersonal goodness. It is fully compatible with the re-interpreted Intuition that adding good lives (even such that are barely good) always makes the world better, just as adding bad lives always makes it worse. This also means that the re-interpreted Intuition of Neutrality is compatible with the Principle of Personal Good that is no longer restricted to the comparisons between worlds sharing the same population:

> *Unrestricted Principle of Personal Good*:
>
> Let $I(X)$ and $I(Y)$ be the populations of worlds X and Y, respectively, and let I be the union of $I(X)$ and $I(Y)$.
>
> (i) If X is better than Y for some individuals in I, while it is equally as good as Y for everyone else in I, then X is better than Y.

[15] This is a context-independent notion of a good life. Ordinarily, when we say that someone's life is good, we implicitly compare it with typical lives that people have in a given social context. Here, though, I aim at a minimal standard of a life's goodness – one that does not vary with social circumstances.

> (ii) If *X* and *Y* are equally as good for everyone in *I*, then *X* and *Y* are equally good.

Until further notice I continue to assume that wellbeing levels are linearly ordered, from higher to lower. (This assumption will be given up in the next section, though.) We can thus think of the wellbeing levels as ordered along a vertical axis, with the personally neutral range located below the levels of good lives and above the levels of bad lives. I can no longer assume, though, that wellbeing is measured on a ratio scale: There is no longer a non-arbitrary zero level of wellbeing. A non-arbitrary zero level could be defined as the level of a life that is personally equally as good as non-existence. But, as we have seen, on the current picture no lives are like that. Good and bad lives are, respectively, personally better and worse than non-existence, while personally neutral lives are all personally incommensurable with non-existence. Thus, on this new picture, measurement of wellbeing on a ratio scale is no longer available. Still, we can continue to assume that wellbeing is cardinally measurable: Wellbeing levels can be represented on an interval scale, with only the unit and the zero point being arbitrary. Indeed, we have available a scale of wellbeing that is somewhat stronger than the interval scale but still weaker than the ratio scale: While the zero level of wellbeing is arbitrarily chosen, this arbitrariness has limits. It is reasonable to require that the zero level should be chosen from the personally neutral range. The reason is that for each wellbeing level in this range, and only for those levels, it is permissible to have preferences that equate that level with non-existence. To put it more precisely: If the life at such level is personally on a par with non-existence, then, as we have seen above, it should be permissible, for the sake of a person who might have this life, to be indifferent between her having this life and her non-existence.[16] In the preference ordering that equates this life with non-existence, its level can therefore be taken as the zero point. In this sense, the levels of personally neutral lives, and only those levels, are permissible candidates for the zero point of the scale.

Indeed, each such permissible scale, with the zero point chosen from the personally neutral range, might be thought of as a *ratio scale* for preferential assessment of lives. This ratio scale specifies a particular permissible configuration of preference strengths for or against different lives, where these preferences are thought of as being held for the sake of a person who could have the lives in question. Thus,

[16] Strictly speaking, each level in the personally neutral range will be on a par with non-existence only if this range forms an *open* interval. If this interval is closed, lives at the upper bound and at the lowr bound of the range still are incommensurable with non-existence, but they aren't on a par with it. It is not permissible to prefer a life at the upper bound of the range to non-existence, nor to prefer non-existence to a life at the lower bound. However, even for each of these boundary levels it is permissible to have preferences that rank lives at this level equally with non-existence.

instead of a unique ratio scale of wellbeing, we now have a *set*, call it *S*, of permissible preferential ratio scales to work with.

This allows us to re-interpret and reformulate *neutral-range utilitarianism* (NRU). Consider any scale *s* in *S*. Let *I*(*A*) stand for the set of individuals that exist in a world *A* and, for any individual *i* in *I*, let *s*(*i*, *A*) be the measure of preference regarding *i*'s life in *A*, as measured on scale *s*. *s*(*i*, *A*) specifies the extent to which *i*'s life in *A* is preferred or dispreferred, as the case may be, for *i*'s own sake, as compared with her non-existence. We can then determine the total sum of the *s*-values of all lives in *A*: $\sum_{i \in I(A)} s(i, A)$. The higher this sum is, the higher is the position of A in the preference ordering of worlds induced by s. The NRU-betterness relation on the set of possible worlds can now be defined in terms of these orderings: World A is better than world B iff A is ranked higher than B in all permissible preference orderings of worlds, i.e. in all orderings induced by the different scales in S.

What impact does the re-interpretation of the neutral range have on NRU's intuitive appeal? The Sadistic Conclusion is, of course, still avoided: Adding bad lives to the world always detracts from its utilitarian value. Indeed, the Weak Sadistic Conclusion is now avoided as well: Any world whose inhabitants have good lives is better than a world inhabited by people with bad lives, worth not living. But, on the other hand, on this re-interpretation re-instates the Repugnant Conclusion: On the re-interpreted account, good lives, even if they are only barely good, add to the value of the world and these additions do not diminish in value as more and more such lives are being added. Therefore, for any world in which everyone's life is excellent, there will be a better world, with a much larger population, in which everyone's life is still good but only barely so.

However, in Rabinowicz (2009) I pointed out that the repugnancy of this re-instated Repugnant Conclusion now is assuaged, if not altogether removed. What in my view made the original Repugnant Conclusion intuitively repugnant was the *short distance* between barely good lives and lives that are positively bad, if only barely so. The former were supposed to be only marginally better than the latter. But, on the re-interpreted conception of the neutral range, the distance between barely good and barely bad lives might well be quite considerable. As I put it: "Lives that are worth living, however modest, cannot be only marginally better than lives that are worth not living, i.e. that are worse than non-existence, if these two kinds of lives are separated by a personal neutral range of non-negligible size." (ibid., p. 406) Indeed, on this picture, barely good lives might be considerably better than drab lives of 'muzak-and-potatoes' variety that Parfit found so unappealing. It is plausible to think that such drab lives are personally neutral rather than positively worth living; plausibly, they are neither better nor worse for us than non-existence.

But then the Repugnant Conclusion does not seem to be so repugnant anymore.[17]

In the next section, however, we shall see that this reassuring diagnosis might have been premature.

Before concluding this section, I should mention that there is another problem with NRU, which I am not going to discuss in this paper. I have in mind what Broome (2004) calls the 'greediness' of neutrality. Suppose that the neutral range contains both level m and levels lower than m-k, for some $k > 0$. As is easily seen, NRU implies that adding to a world A a person with a life at level m, while at the same time decreasing by k units the wellbeing of one of the originally existing persons results in a world, B, that is incommensurable with A. For while choosing m as the benchmark gives rise to a preference ordering of worlds in which A is ranked above B, setting the benchmark at a wellbeing level lower than m-k yields a preference ordering in which B is ranked above A.[18] Thus, as Broome (2004, p. 170) puts it: "Incommensurateness [...] is a sort of greedy neutrality, which is capable of swallowing up badness and goodness and neutralizing it. This is implausible [...]"

This problem is exacerbated given my re-interpretation of the neutral range. On this re-interpretation, the life of the person added in B is not even good for her (though it is not bad for her either). But B still is not worse than A, even though it makes life positively worse for one of the originally existing people. This bad thing about B, as compared with A, is swallowed up by B's incommensurability with A. Effects like this might seem implausible, but I am inclined to believe that they should be accepted. They are just part of the package that comes with the existence of a neutrality range. For a critical discussion of the greediness objection, see Rabinowicz (2009).

4. Complicating the picture—incommensurable lives

Let us define a *strictly neutral* life as a life that is personally equally as good as non-existence. Obviously, all strictly neutral lives are neutral, i.e., neither personally better nor personally worse than non-existence, but the converse doesn't hold. As we have seen in the preceding section, if a neutral life is personally better or worse

[17] An essentially similar treatment of the Repugnant Conclusion is defended by Gustafsson (2016, who re-interprets the neutral range in the way as I have done in Rabinowicz (2009).

[18] The same conclusion – that B is incommensurable with A – can also be established using weaker premises, if one proceeds less directly and considers some other possible worlds as well, along with A and B. It is then sufficient to rely on the utilitarian balancing of gains and losses only in comparisons between worlds that share the same population. Furthermore, there is then no need to assume my account of incommensurability, in terms of divergent permissible preference orderings. (Cf. Broome, 2004, p. 170, and Rabinowicz 2009.)

than some other neutral life, then it cannot be strictly neutral: The hypothesis that it is personally equally as good as non-existence cannot be upheld, on pain of a contradiction. We can refer to such lives as *weakly neutral*. A life is weakly neutral iff it is incommensurable with non-existence in its personal value.

If strictly neutral lives can exist, along with weakly neutral lives, we must give up the assumption that all lives' wellbeing levels are linearly ordered. The wellbeing level of a strictly neutral life cannot be fitted into such an ordering. Unlike weakly neutral lives, a strictly neutral life cannot be personally better or worse than any other neutral life. And it cannot be personally equally as good as any weakly neutral life: By the transitivity and symmetry of equal goodness, it can only be equally as good as other lives that are equally as good as non-existence, i.e., it can only have this relation to other strictly neutral lives.

Postulating the possibility of strictly neutral lives along with lives that are weakly neutral thus requires that we allow for the existence of lives that are mutually incommensurable in their personal value. But, on reflection, this is something that we should allow for anyway. Surely, it is implausible to insist that for any two lives with different wellbeing levels, the wellbeing level of one must be higher or lower than that of the other life. Life wellbeing is a *many-dimensional* concept: Specifying its level requires characterizing a life with respect to several relevant dimensions. One life might be better than another in some respects, and worse in other respects. At the same time, different weight assignments to the relevant respects of comparison might be permissible and the all-things considered preference ordering of lives will depend on how these respects are weighed against each other. Consequently, a life L might be preferred to another life L', for the sake of a person who might live one of these lives, given one permissible weight assignment, and dispreferred given another. This would imply that *L* and *L'* are incommensurable in their personal value, or – what amounts to the same – that *L* and *L'* have different wellbeing levels even though none of these levels is higher than the other. Thus, we can no longer assume that wellbeing levels of lives are linearly ordered.[19]

Nevertheless, while we should for this reason be willing to accept that lives might well be mutually incommensurable in personal value, we might still wonder whether to allow for the existence of strictly neutral lives. Gustafsson (2016, section 5) is unwilling to admit this possibility. He uses a different label for lives of this

[19] That some lives (namely, the ones that are weakly neutral) can be incommensurable with non-existence in their personal value has a similar explanation, by the way. A life typically has both desirable and undesirable components and its assessment on balance, in comparison with non-existence, might depend on how these components are weighed against each other. Since different weight assignments will normally be permissible, they might give rise to opposing permissible preferences regarding the life in questions: It might be permissible, for the sake of the person who could live this life, to prefer it, all things considered, to her non-existence, but it might also be permissible to have the opposite preference, for her sake.

kind[20] and he does not define them by comparing their personal value with non-existence – indeed, he questions the possibility of such comparisons. But he takes such lives, *if* they exist, to satisfy the following three principles:

> For any strictly neutral life *L* and for every life *L'*,
>
> (i) *L'* is good iff it is (personally) better than *L*,
>
> (ii) *L'* is bad iff it is (personally) worse than *L*,
>
> (iii) *L'* strictly neutral iff it is (personally) equally as good as *L*.[21]

This squares with the definition I have proposed: As is easily seen, the definition of a strictly neutral life as one that is personally equally as good as non-existence entails (i), (ii) and (iii), assuming that the relation of equal personal goodness is transitive and that personal betterness is transitive across this relation.

One potential candidate for a strictly neutral life that Gustafsson (2016) considers, but finally rejects, is a life devoid of any good or bad components.[22] An example of a life of this kind would, I take it, be a life in a permanent state of unconsciousness.[23] Gustafsson denies, however, that such a life is possible for a person. One is not a person if one is never conscious. Being a person requires having some psychological features and being in some psychological states, but no such states and features can be present if consciousness is permanently absent.

I don't think this shows what it is supposed to show. It doesn't show that a person could not live such a life. It is true that I wouldn't be a person if my whole life were spent in a coma. But it still is true that I – a person – could have had such a life (in which I wouldn't be a person). It is a possible life for me, a person. While it is a popular view that being a person is an essential property of persons, I think this view is mistaken: A person could have had a life in which she wouldn't be a person. But

[20] He simply calls them "neutral" and refers to lives I call weakly neutral as "blank" (or "undistinguished", in later versions of his draft). In my comments on Gustafsson, I will however, continue to use my own terminology, to avoid confusion.
[21] Note that if a life *L* is weakly neutral, it need not satisfy (i) and (ii): a life *L'* might be better than *L* without being good and it might be worse than *L* without being bad. Rather than good or bad, respectively, *L'* might itself be weakly neutral.
[22] Indeed, if such a life is to be strictly neutral, it should also lack weakly neutral components, i.e., it should lack components that it is permissible to prefer (for the sake of the person who could lead this life), but also permissible to disprefer.
[23] In Broome's terminology, a life without any good or bad experiences is a 'blank life', and a life spent in a coma is an example of a blank life. (Cf. Broome 2004, pp. 208f.)

then for me, a person, my counterfactual life as a non-person, in a permanent state of unconsciousness, might be equal in value with my (equally counterfactual) non-existence. Consequently, this life might be strictly neutral.

Whether a life in a coma in fact is strictly neutral might of course be questioned. On some views, such a life would be worse for me than non-existence, due to considerations relating to human dignity. Perhaps a better candidate for a strictly neutral life could be found. Or perhaps not. The claim that strictly neutral lives could exist is logically coherent, but it is not clear to me whether such lives really are possible. Perhaps it might be argued that, for any possible life, it is at least permissible to prefer it or permissible to disprefer it (or both) to non-existence, for the sake of a person who could live that life. In such a case, no possible life would be strictly neutral. Let me now, however, consider what it would mean if strictly neutral lives could exist.

A neutral-range utilitarian might find the possibility of such lives quite worrying, as it implies, as we have seen, that lives might be mutually incommensurable in personal value. One might think that such incommensurabilities would make NRU meaningless: It would no longer be possible to determine the total value of a world's population for different choices of benchmarks from the neutral-range.[24]

But, as I have pointed out, even if there were no strictly neutral lives to reckon with, we would still have to accept that some lives can be mutually incommensurable in personal value. If it were true that such incommensurabilities make NRU meaningless, then this theory would not be worth serious consideration. I think, though, that this worry can be put to rest: NRU can accommodate incommensurabilities between lives. Let me explain how it can be done.

On the interpretation of NRU I have given in the preceding section, this theory assumes that different permissible preference orderings of worlds are generated by different permissible preferential ratio scales for assessing lives – scales whose zero points are drawn from the neutral range. If life wellbeing is cardinally measurable, as we have previously assumed, then these scales only differ in their choices of zero from the neutral range, i.e., in their choices of the neutral wellbeing level such that lives on that level are equi-preferred with non-existence on a given scale. (The scales might also differ in their choices of the unit of measurement, but this choice doesn't natter when we do utilitarian calculation in order to compare worlds with each

[24] This is indeed the reason why Gustafsson (2016), who proposes an ethical theory that is very much like NRU (he calls it 'critical-range utilitarianism') and who like myself works with the personalized neutral range, is so opposed to strictly neutral lives. However, the reasons he offers for rejecting such lives change in the later versions of his paper. There, the meaninglessness worry no longer is mentioned. And even in the original version he eventually suggests a version of critical-range utilitarianism that does allow for incommensurable lives. On that theory, utilitarian aggregation of wellbeing in a population is done at the level of life moments instead of the whole lives. But he still assumes that, at the level of life moments, there are no incommensurabilities.

other. And since world comparisons are done independently for each scale, it doesn't matter either if different scales use different units of measurement.) However, if incommensurabilities between lives are allowed, the assumption of cardinal measurability of wellbeing must be given up.[25] Not only will the choice of zero differ, as before, between different permissible preferential ratio scales, but now some such scales will also differ in their *ordering* of lives: They will differ in how they rank incommensurable lives against each other. For example, if lives L and L' are personally on a par, then some permissible scales will rank L above L', while others will rank it below.[26] The set S of permissible ratio scales will thus be much more varied than it was previously assumed. Class K of permissible preference orderings of worlds will however still consist of the orderings that are generated by the scales in S, as before. For any scale s in S, the position of a world A in the world ordering P_s induced by s is determined by the sum $\sum_{i \in I(A)} s(i, A)$. The higher this sum is, the higher is A's position in P_s. The value relations between worlds are then determined by this class K of permissible preference orderings of worlds, in the standard way: world A is better than world B iff it is ranked above B in all permissible orderings; A is equally as good as B iff it is equal-ranked with B in every permissible ordering; A is incommensurable with B iff none of these worlds is better than the other nor are they equally as good.[27]

Clearly, on this version of NRU, there will be many incommensurabilities between worlds, many more than if all lives were commensurable in their personal value. But this does not, by itself, undermine a utilitarian axiology.

Nor does this account undermine the basic welfarist principle that all (impersonal) value comparisons between worlds are determined by the wellbeing levels of the individuals existing in those worlds. However, the wellbeing level of a life no longer can be represented by a single number. Instead, we can now represent it as a function that assigns a numerical value to every scale in S. These values specify how highly a life at this wellbeing level is assessed on each scale. If all these values are positive, the wellbeing level is positive; if they are all negative, the wellbeing level is

[25] What is not given up is cardinal measurability of the *strength of preference* for different lives, according to each permissible preferential scale. But a life's wellbeing, which is characterizable by the position of this life on all the different permissible preferential scales, is no longer cardinally measurable. Nevertheless, lives' wellbeing levels can still be given a numerical represntation, as will be shown below.

[26] Note, though, that if strictly neutral lives can exist, they will be placed in all those scales at the zero level. On any such scale, the zero level is the level of lives that are equi-preferred, for the sake of a person who could have that life, with her non-existence. Thus, in one way or another, each scale will commensurate strictly neutral lives with lives that are weakly neutral.

[27] Strictly speaking, class K will also contain incomplete world orderings. But if any such permissible incomplete ordering is the intersection of some set of permissible complete orderings, then adding incomplete orderings to K will not affect the extensions of the relevant value relations between worlds: the relations of betterness, equal goodness, incommensurability, and parity.

negative. If neither holds, the wellbeing level is neutral. In particular, for a strictly neutral life, all the values are zero. Two lives have the same wellbeing level if they are characterized by the same function of this kind. One wellbeing level, *w*, is higher than another, *w'*, iff for all scales *s* in *S*, the *s*-value of the function that represents *w* is higher than the *s*-value of the function that represents *w'*. In other words, the wellbeing level of one life, *L*, is higher than that of another life, *L'*, iff on every scale in *S*, *L* is ranked above *L'*. We can also define what it means that one wellbeing level is at least as high as the other: This is the case iff for all scales *s* in *S*, the *s*-value for the former level is at least as high as the *s*-value for the latter level. (Note that, given this definition, a life might be at a wellbeing level that is at least as high as another life without the level of the former life being higher than or equal to that of the latter life.) As is easy to see, wellbeing levels are partially ordered by this at-least-as high-as relation. In other words, the relation in question is reflexive, transitive and anti-symmetric. (Anti-symmetry means that if level *w* is at least as high as level *w'* and *w'* is at least as high as *w*, then *w* is identical to *w'*.)

Thus, the worry that NRU becomes meaningless if strictly neutral lives are allowed and, more generally, if lives are allowed to be incommensurable, is unjustified. It is still possible to define lives' wellbeing levels and carry out utilitarian calculations in order to determine the value of each world. This value is a function that for each scale *s* in *S* specifies the total sum of the *s*-values of the lives of the individuals that exist in the world in question.

It is another issue, however, whether NRU, so interpreted, is an intuitively appealing theory. This might be questioned. To explain why, let me first point out that allowing for strictly neutral lives leads to rather unexpected and unwelcome consequences. Suppose, for definiteness, that a life *L*, which is wholly spent in a coma, is strictly neutral, and consider a life L^+ that is slightly personally better than *L*. L^+ might, for example, be a life that also is mainly spent in a coma, apart from a very short period during which its subject is conscious and experiences a moderate sensory pleasure (and nothing else). Since L^+ is slightly (personally) better than *L*, which is (personally) equally as good as non-existence, it follows that L^+ is slightly (personally) better than non-existence. Which implies that L^+ is a good life, though only barely so. In the same way, we can think of a barely bad life, L^-, which is slightly (personally) worse than *L* and thus slightly (personally) worse than non-existence. L^- might be a life mainly spent in a coma, apart from a very short period during which its subject is conscious and experiences a moderate pain (and nothing else).

Now, note that the value distance between L^+ and L^- is short: L^+ is only marginally better than L^-. Thus, we now have to accept that there are good lives that are marginally better than bad lives.

But how can this be? Didn't we previously show that wellbeing levels of good lives

are separated from the wellbeing levels of bad lives by the neutral range, which might be quite extended?

That was then, though, before we gave up the assumption that wellbeing levels are linearly ordered. On the new picture, things look very differently. Indeed, as we already know, a strictly neutral life L, if such a life can exist, is incommensurable in its personal value with all lives that are weakly neutral. And it is arguable that the same applies to the barely good life L^+ and the barely bad life L^-. As has been noted by Joseph Raz, it is a "mark of incommensurability" that if this relation obtains between two items, then a small improvement or a small worsening of one of the items need not (and typically will not) remove their incommensurability (see Raz 1986, p. 26). Thus, since incommensurability obtains between the strictly neutral life L and all lives that are weakly neutral, it might well still obtain when L is replaced by L^+ or L^-.

This means that on the new picture a good life such as L^+, and a bad life such as L^- might be incommensurable with a neutral life – a life that is neither bad nor good. Indeed, they might both be incommensurable with all neutral lives that are weakly neutral. Some good lives (such as L^+) need no longer be better than all lives that aren't good and some bad lives (such as L^-) need no longer be worse that all lives that aren't bad. This might be surprising, but it is an implication that we now must accept.

Note also that this implication does not strictly speaking require the possibility of strictly neutral lives. What it does require is the possibility of lives *close to strict neutrality*: lives that are only slightly better or slightly worse than non-existence. Since weakly neutral lives are incommensurable with non-existence, it is to be expected that they will also be incommensurable with lives such as L^+ and L^-.

Thus, the value distance between a barely good life and a barely bad life might be very short. The former might be only marginally better than the latter.[28] L^+ and L^- provide a case in point.[29] But this means that the Repugnant Conclusion which follows from the re-interpreted NRU regains its original repugnance. Just as on the standard total-sum utilitarianism, we are now driven to the conclusion that every

[28] In private communication, John Broome has suggested the a situation like this will arise if "there are pairs of lives such that one is better than non-existence and one is worse than non-existence, and they are very similar to each other."

[29] But if we no longer assume that lifetime wellbeing is cardinally measurable, how can we say that L^+ is only *marginally* better than L^-? It might be objected that in the absence of cardinal measurement, differences in lifetime wellbeing cannot be judged to be small or large. I don't think, though, that this objection is compelling. Pockets of cardinality – areas in which size estimates of wellbeing differences are meaningful – might still exist, even though cardinal measurement no longer is possible in all cases – in all comparisons of lives' wellbeing levels. Those are the areas in which all the permissible ratio scales in S agree in their assessments of differences between life levels. The difference between two wellbeing levels is small if it is small on each of the permissible scales.

world in which everyone has an excellent life is worse that some world in which everyone's life is only marginally better than a life worth not living.

This is bad news for neutral-range utilitarianism. But its adherents might be well-advised to stand fast and hold on to their view. The truly repugnant Repugnant Conclusion crucially depends on the possibility of lives close to strict neutrality. But this possibility, just as the possibility of strictly neutral lives themselves, has *not* been positively established. Until it has been done, which might never happen, NRU remains unchallenged. Its adherents might persist in denying that strictly neutral lives and lives close to strict neutality really are possible.[30]

However, just as a thought experiment, suppose someone proves to us that such lives indeed are possible. Then NRU will have to be given up unless we are willing to accept the repugnant Repugnant Conclusion. But, if we give up NRU, are there some claims from the preceding discussion that we still can we retain?

I think we can retain the Intuition of Neutrality itself, in its re-interpreted, personalized version:

> *Adding personally neutral lives to the world is impersonally neutral: It does not make the world either better or worse.*

We can also retain the insight that:

> *Lives can be incommensurable with each other in their personal value.*

In other words, wellbeing levels are not linearly ordered.[31]

Despite this acceptance of incommensurabilities between lives, we can, if we wish, hold on to some of the central tenets of welfarism:

(i) *The Unrestricted Principle of Personal Good*,

and the assumption that:

[30] A mere logical possibility of such lives is not enough to undermine NRU. I am indebted to John Broome for a discussion of this point (in private communication).

[31] This of course also applies to levels in the neutral range, and it would apply to them even if there were no strictly neutral lives. Lives that are weakly neutral can be mutually incommensurable. But still, it is reasonable to suppose that for every such level m there is some neutral level that is higher or lower than m. This means that we can continue to uphold the assumption of the proof provided in Section 1. As we remember, that proof was meant to establish that a world in which a person is added at some such level m must be incommensurable with the original world.

> (ii) *Value comparisons between worlds are determined by the wellbeing levels of individuals who exist in those worlds.*[32]

Furthermore, if lives close to strict neutrality are possible, we can draw the rather surprising lesson that:

> *A good life need not be better than a neutral life and a bad life need not be worse than a neutral life.*

A life that is better (worse) than non-existence is good (bad) but it need not be better (worse) than a life that is incommensurable with non-existence. It might be incommensurable with such a life.

This lesson generalizes: It is potentially applicable to all analyses of 'good' and of other monadic value predicates in terms of value comparisons with some *standard*. In this paper, when defining monadic value predicates, the items we have targeted have been lives and the standard we have chosen has been non-existence. But we can consider other items and choose other standards. Thus, suppose we consider some domain of items, pick out a standard σ and adopt the following definitions: For all items x in this domain,

> x is good iff x is better than σ,
>
> x is bad iff x is worse than σ,
>
> x is neutral iff x is neither better nor worse than σ,
>
> x is strictly neutral iff x is equally as good as σ.

Then it does not follow that a good x must, by logical or analytical necessity, be better than a neutral y, or that a bad x must be worse than such a y.[33] These seemingly very plausible entailments do not obtain if y, while neutral, is not strictly so, i.e. if y is incommensurable with the adopted standard. In such cases, a good x will typically not be better than y if x is only slightly better than the standard. Likewise, a bad x

[32] But, in view of the incommensurabilities between lives, wellbeing levels no longer can be assumed to be representable by single numbers. Their numerical representation has to be more complicated, as we have seen above.

[33] This has already been noted in Gustafsson (2016).

will typically not be worse than such *y* if *x* is only slightly worse than the standard.

Thus, some of the things we have learned have implications that go beyond population axiology.[34]

References

Arrhenius, G. (2000) "An Impossibility Theorem for Welfarist Axiology," *Economics and Philosophy* 16: 247–266.

Arrhenius, G. (2011) "The Impossibility of a Satisfactory Population Ethics", in E. Dzhafarov and L. Perry (eds.), *Descriptive and Normative Approaches to Human Behavior*, World Scientific: 51–66.

Arrhenius, G. (2016) "Population Ethics and Different-Number-Based Imprecision", *Theoria* 82: 166–181.

Arrhenius, G, forthcoming, *Population Ethics: The Challenge of Future Generations*, Oxford University Press.

Arrhenius, G., & Rabinowicz, W. (2010) "Better to Be Than Not to Be?" in H. Joas and B. Klein (eds.), *The Benefit of Broad Horizons*, Brill: Leiden, 399–421.

Arrhenius, G., & Rabinowicz, W. (2015) "The Value of Existence," in in I. Hirose and A. Reisner (eds.), *The Oxford Handbook of Value Theory*, Oxford University Press: Oxford, 424–444.

Blackorby, C., Bossert, W., & Donaldson D. J., (1996), "Quasi-orderings and Population Ethics," *Social Choice and Welfare* 13: 129–150.

Blackorby, C. &Donaldson, D. (1984), "Social criteria for evaluating population change". *Journal of Public Economics* 25:13–33.

Broome, J. (1999) *Ethics out of Economics*, Cambridge University Press: Cambridge.

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

Broome, J. (2009) "Reply to Rabinowicz," *Philosophical issues* 19: 412–417.

[34] Early versions of this paper were presented at a conference on formal ethics in York in June 2017 and at a conference on philosophy and economics in Canberra the same year. I am indebted to the participants of these events for several useful suggestions. I also want to thank John Broome, Johan Gustafsson and Gustaf Arrhenius for their helpful written comments and discussion. Broome's challenging comments, in particular, have caused me to modify some of the conclusions of this paper.

Bykvist, K. (2007) "The Benefits of Coming into Existence," *Philosophical Studies* 135: 335–62.

Bykvist, K. (2015) "Being and Well-Being," in I. Hirose and A. Reisner (eds.), *Weighing and Reasoning*, Oxford University Press: Oxford.

Carlson, E., this volume, "On Some impossibility Theorems in Population Ethics".

Darwall, S. (2002), *Welfare and Rational Care*, Princeton University Press: Princeton and Oxford.

Gustafsson, J. (2016), "Population Axiology and the Possibility of a Fourth Category of Absolute Value," draft. A revised version has appeared in *Economics and Philosophy* 36, 2020, pp. 81–110.

Holtug, N. (2001) "On the Value of Coming into Existence," *Journal of Ethics* 5: 361–84.

Johansson, J. (2010) "Being and Betterness," *Utilitas* 22: 285–302.

Narveson, J. (1973) "Moral Problems of Population," *The Monist* 57: 62–86.

Parfit, D. (1991 [1984]) *Reasons and Persons*, Clarendon Press: Oxford.

Rabinowicz, W. (2008), "Value Relations," *Theoria* 74: 18–49.

Rabinowicz, W. (2009) "Broome and the Intuition of Neutrality," *Philosophical Issues* 19: 389–411.

Rabinowicz, W. (2012), "Value Relations Revisited," *Economics and Philosophy* 28: 133–164.

Raz, J. (1986), The Morality of Freedom, Oxford: Clarendon Press.

Rønnow-Rasmussen, T. (2011), *Personal Value*, Oxford University Press: Oxford.

Rønnow-Rasmussen, T. (2018), "Fitting-Attitude Analysis and the Logical Consequence Argument," *The Philosophical Quarterly* 272: 560–579.

Taurek, J. (1977), "Should the Numbers Count?" *Philosophy and Public Affairs* 6: 293–316.

Krister Bykvist[1] & Tim Campbell[2]

Persson's Merely Possible Persons[3]

In many cases, it seems, one possible outcome is worse than another in virtue of the well-being of people who do not exist in both. For example, it seems, creating a very unhappy person makes the world worse, other things being equal. And some would say that we make the world better, other things being equal, by creating a very happy person. It would be easy to justify such claims if it can be better, or worse, for a person to exist than not to exist. But that seems to require that things can be better, or worse, for a person even in a world in which she does not exist, which sounds paradoxical. This paradoxical-sounding claim has been defended by Ingmar Person. He argues that in a world in which a person does not exist, she is a merely possible being – a being that has never existed and never will – and that for such beings it is worse not to exist than to exist with a good life. Furthermore, he argues for this claim from what he claims are "incontestable" premises. We argue that the premises are far from incontestable. The argument, as stated by Persson, has false premises and is invalid. We can reconstruct the argument to make it valid, but this still leaves us with some clearly contestable premises. Finally, we will argue that it is possible to make sense of our obligations to future generations without letting merely possible beings into the moral club.

[1] Institute for Futures Studies & Department of Philosophy, Stockholm University, krister.bykvist@iffs.se.
[2] Institute for Futures Studies, timothy.campbell@iffs.se.
[3] Thanks to Ingmar Persson for helpful discussion and comments on earlier drafts of this paper. Financial support by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences) is gratefully acknowledged.

I

Many of our choices effect who will exist in the future, not just the obvious choice of having a child but also the choice of giving priority to young people over old when saving lives or offering generous state-funded parental leave. In at least some of these cases we want to say that one outcome is worse than another in virtue of the wellbeing of people who do not exist in both. For example, we want to say that creating a very unhappy person makes the world worse, other things being equal. Some would also say that we make the world better, other things being equal, by creating a very happy person (or a sufficiently large number of such persons). It would be easy to justify these verdicts if it can be better, or worse, for a person to exist than not to exist. But can it really be better, or worse, for a person to exist than not to exist? That seems to require that things can be better, or worse, for a person even in a world in which she does not exist, which sounds paradoxical.

This paradoxical-sounding claim is defended in Ingmar Person's latest book *Inclusive Ethics*.[4] More specifically, he argues that in a world in which a person does not exist, she is a merely possible being – a being that has never existed and never will – and that for such beings it is worse not to exist than to exist with a good life. Furthermore, he argues for this claim from what he claims are "incontestable" premises. We shall argue that the premises are far from incontestable. In fact, the argument, as it stated, has obviously false premises and is also invalid. It is possible to reconstruct the argument so that it becomes valid, but this still leaves us with some clearly contestable premises. Finally, we will argue that it is possible to make sense of our obligations to future generations without letting merely possible beings into the moral club.

II

Persson's master argument is as follows:[5]

> (1) For a being who has never existed, nothing is either intrinsically good or bad.
>
> (2) The fact that nothing is either intrinsically good or bad for a being is worse for it than the fact that things are overall intrinsically good for it.

Therefore,

[4] Persson (2017).
[5] Persson (2017: 61).

> (3) Not existing is worse for a being than existing with a life in which things are overall intrinsically good for it.

First, an important clarification: (1) talks about a being who *has never* existed. This does not rule out that it *will* exist in the future. But Person wants to say that non-existence can be worse for *merely possible* beings, beings that not only never *have* existed but also never *will* exist. So, readers should interpret (1) and (3) as talking about merely possible beings.

One problem with the argument concerns (2). Persson insists that (2) is "incontestable,"[6] but as stated it is clearly false. This is because the two different facts mentioned in (2) are *incompatible*. Necessarily, if it is a fact that nothing is intrinsically good or bad for a certain being, then it is *not* a fact that things are overall intrinsically good for this being. Hence, it can't be true that *the fact* that nothing is intrinsically good or bad for a being is worse for it than *the fact* that things are overall intrinsically good for it.[7]

Another problem is that the facts compared in (2) are *evaluative* facts, facts about whether things are good or bad for a being. But it is not clear whether it can be true that one evaluative fact is worse than another for a being.

Third, the argument, as it stands, is invalid—(3) does not follow from the conjunction of (1) and (2). According to (2), a certain *evaluative fact* is worse for a being than a certain other *evaluative fact*. But 'not existing' in (3) refers to the *non-evaluative* state of affairs (or fact) of a being not existing.

III

The following reconstruction of Persson's argument, where evaluative fact-talk is replaced by non-evaluative state-of-affairs-talk throughout, avoids the problems of the originally-stated argument:

> (4) For a merely possible being, non-existence is neither overall good nor overall bad.

> (5) Any state of affairs which is neither overall good nor overall bad for a being is overall worse for it than a state of affairs that is overall good for it.

[6] Persson (2017: 61).
[7] This is an instance of the 'relata problem'. See, for example, Arrhenius & Rabinowicz, (2010: especially 404-408), Arrhenius & Rabinowicz (2014: 432), Bykvist (2015: 90-91), Holtug (2010: 140-141).

Therefore,

> (6) For a merely possible being, non-existence is overall worse than an existence that is overall good for it.

This reconstruction of Persson's argument is logically valid, assuming 'existence' and 'non-existence' pick out states of affairs. Moreover, it avoids incoherent evaluative comparisons of incompatible evaluative facts. Finally, we have made clear that the relevant comparisons concern the overall values of states of affairs (existence or non-existence), i.e., the total intrinsic value for the being of all the things that the states of affairs would realize, if they obtained.[8] It should be added that the reason why (4) is true is that everything lacks positive and negative intrinsic value for such a being.

But the argument is still problematic. Take premise (5) first. This premise states that something that is overall good for a being is better for a being than something that *lacks* overall positive and negative value for it. But in order for something to better for a being than something else both things need to have value for the being. The same holds for any other comparative notion. For one thing to be taller than another, both things need to have length. For one thing to be heavier than another both things need to have weight. For one thing to have a higher temperature than another, both things need to have temperature. The question is then what value non-existence has for a merely possible being.

Persson states that it has neutral value for the merely possible being. Indeed, he claims that everything is neutral for a merely possible being. However, he characterizes 'neutral value for a being' as something that is neither intrinsically good nor intrinsically bad for a being'.[9]

But this is not a plausible characterization of 'neutral value for a being' because, plausibly, some states of affairs are *undefined* in value for a being. Possible candi-

[8] In conversation, Persson has told us that he accepts this reconstruction, but would prefer (2) to be formulated subjunctively as 'Any state of affairs which would be neither overall good nor overall bad for a being would be overall worse for it than a state of affairs that would overall good for it' just to make clear that states of affairs have value for a being only when they obtain. We shall assume this reading implicitly in the following. He has also told us that he prefers that the premises be stated in terms of intrinsic goodness and badness for a being rather than in terms of the overall goodness and badness that states of affairs have for a being. This is because he thinks that non-existence could be extrinsically good or bad for a being by excluding an existence that would be good or bad for that being. On our view, the overall goodness or badness that a state of affairs has for a being does not depend on this kind of preventive value.

[9] Persson (2017: 11, 57, 61). Nils Holtug defends a similar definition of 'zero value for a being' in his (2001), but he has since given it up because of the objection raised here. For further debate about the intelligibility of assigning neutral or zero value to non-existence, see Roberts (2003), Johansson (2010), Bykvist (2007), Bykvist (2015: 90-91).

dates are contradictory states of affairs (e.g., that 2 +2 = 5), necessary states of affairs (e.g., that 2 + 2 = 4), evaluative states of affairs (e.g., that happiness is good for people), and states of affairs that concern other people's wellbeing (e.g., that a stranger is unhappy).[10]

Furthermore, being neutral for one is different from being neither good nor bad for one. If something is neutral for one, then it has a certain *value* that is neither positive nor negative, but which can be compared to positive and negative values. This is analogous to

> (a) having zero temperature, which is to have certain temperature that is neither positive nor negative, and thus different from lacking positive and negative temperature just because one lacks any temperature, or
>
> (b) having a weight that is neither heavy nor light, which is different from being neither heavy nor light just because one lacks any weight, or
>
> (c) having a height that is neither tall nor short, which is different from being neither tall nor short just because one lacks any height.[11]

Persson could deny that things must have value to stand in value relations, or deny that neutral value is a value in its own right. But such claims are highly contestable, and we need arguments for why value comparisons differ so radically from other comparisons. No such argument is provided in the book.

Premise (4) is also far from incontestable. Indeed, the claim that there are merely possible beings is one of the most contestable claims in modal metaphysics.[12] Persson claims not only that there are beings that do not exist, but also that things are neutral for them and, in virtue of premise (5), that things can be better or worse for them.

Consider the existence claim first. On its face, it seems incoherent, since it seems to assert the *existence* of *non-existent* beings. But Persson denies the charge of incoherence because he thinks that there are different senses of 'exist' and 'there are':

---

[10] On the distinction between being neutral for and having undefined value for, see Bradley (2009: 98-104), Luper (2007).

[11] These examples are relevant not only to the debate about the value of existence, but to the more general issue of how to understand the nature of properties. For more on this general issue, see Balashov (1999).

[12] For a critique of the view that there are merely possible beings, see Stalnaker (2012: especially chs. 1 and 2).

> In one sense of 'exist', it is true that there are merely possible beings because this follows from the clearly true claim that it is possible that some beings will begin to exist in the future. This is the sense in which I believe there to be merely possible and, thus, non-existent beings. I cannot provide a philosophically adequate explication of this sense, but there are many commonsensical claims to which we can permissibly help ourselves, though we cannot accurately expound their sense philosophically. (2017: 60-61)

Persson's claim that

> A. there are merely possible beings

follows from

> B. it is possible that some beings will begin to exist in the future.[13]

is puzzling. B states that in some possible world some beings will begin to exist in the future. Why would it follow (logically) from this claim that there are and thus exist, *in a different sense*, merely possible beings *here in the actual world*? Compare: it is possible that the pope has two children in the future, and thus possible that two children of the pope will begin to exist in the future. But why would it follow from this that there exist, *in a different sense*, merely possible beings *here in the actual world*? Even if we grant Persson that there are different senses of existence, we still need to know how to derive A from B. We could derive it, if we accepted the Barcan formula as a bridge principle:

> If it is possible that there is a being that is F, then there is a being such that it is possible that it is F.

For then we could make use of the following instance of the Barcan formula, (where 'there are' is equated with 'there exist', in which 'exist' has a different meaning from 'exist' in 'begin to exist'):

[13] We take B to say that it is possible that it will be the case that some beings begin to exist, not as saying that it is possible that there are now some beings that will begin to exist, since the latter claim commits one to the possibility of merely future beings, beings that do not exist yet but will exist in the future, a commitment which is almost as contestable as the commitment to merely possible beings.

> If it is possible that there are beings that will begin to exist, then there are beings that are such that it is possible that they begin to exist.

and from B derive that there are beings that are such that it is possible that they begin to exist, which entails A, assuming that they in fact will not begin to exist. But the Barcan formula is far from incontestable, and Persson claims (in correspondence) that this is not what he had in mind.[14] So, we are at a loss as to how he can incontestably derive A from B.

Even if we could somehow establish that A follows from B, we need to know what the *relation* is between merely possible beings in one world and beings that begin to exist in some alternative possible world. Either they are identical, or they are not. Both options are problematic.

If they are identical, then we must say that a merely possible being which is not concrete (not located in space or time and lacks causal powers) could have been concrete in the sense of beginning to exist, i.e., existing in time. But being non-concrete is surely a very good candidate for a property that is essential to its bearers: if something is non-concrete, then it is essentially non-concrete. Similarly, if something is concrete, it is essentially concrete.

If they are not identical, then we will have problems understanding how Persson's view applies to existing *flesh and blood people*. Consider for example, a child born with a painful and fatal condition who lived in agony for a few months before dying. Call this child Tommy, and suppose that he lived a horrible life. Persson seems committed to saying that things would have been better for Tommy if he had not existed. As we point out in the last section, according to Persson all reasons are comparative, which he takes to mean that in order to say that we have reasons of beneficence not to create someone we need to show that not creating the person would be *worse* for the person. Now, we do want to say about Tommy that we did have at least some reason not to create him (but perhaps not overall reason). But if we had not created Tommy, he would not have been around in any sense, since being concrete is essential to Tommy. Since Tommy would not have been around, things would not have been worse for him, for 'worse for' is a relation that requires a subject for whom things are worse. Now, it is true that we are assuming that there is a *distinct* merely possible being in the Tommy-less world. But even if things are worse for it, this does not establish that things are also worse for *Tommy* in this world, for we have assumed that the merely possible being is not identical to Tommy.[15]

[14] See Williamson (2013: ch. 2) for a discussion of the Barcan formula.

[15] In conversation, Persson seems to go for this horn of the dilemma. He compares merely possible beings to abstract properties and concrete beings with property bearers. Merely possible beings are

Now let us turn to the value claim: that things can be neutral for, or better or worse for, merely possible beings. It is far from obvious that things can be better or worse for merely possible beings (when they are merely possible beings). After all, a merely possible being is not a concrete being, an animal, a conscious being, a human, a male or female, or a parent or non-parent. At most, it is a *merely possible* concrete being, a *merely possible* animal, and so on. Why think that a being that merely possibly exemplifies any of these features can nevertheless stand in value relations? Why not think instead that standing in a value relation requires being concrete in some way, for example, having a mind, body, being in space or time, or having causal power? We need an argument for the controversial claim that things can be better or worse for merely possible—and hence non-concrete—beings.

Persson offers an analogy to support this claim. The analogy involves a comparison between a non-existent being and a certain type of existing being—an anencephalic infant—that lacks the consciousness-generating parts of the brain and so never becomes conscious:

> It seems indisputable that, given that existing without consciousness, like anencephalic infants do, is neither intrinsically good nor bad for them, this is worse for them than having consciousness and leading a life in which things are predominantly good for them. But we have seen that non-conscious beings are like non-existent beings in that nothing is either intrinsically good or bad ... for them. Therefore, non-existence is worse for a being than a predominantly intrinsically good existence, just as its existence is worse for a non-conscious being than a good existence. (2017: 62)

According to Persson, the crucial similarity between the anencephalic infant and the merely possible being is that both lack consciousness. This is supposed to show that nothing is either intrinsically good or bad for them, and hence that the lack of conscious existence is neutral for them. The analogy is contestable, however, for there is a clear difference between the two beings. Although both lack consciousness, the anencephalic infant, unlike the merely possible being, is a concrete being and an animal. The merely possible being is only *merely possibly* concrete and *merely possibly* an animal. For this reason, the anencephalic infant is a better candidate for a being for which things can be neutral. Remember that being neutral for is not just a lack of the relations being good for and being bad for; it is an evaluative relation in its own right. Note also that the analogy completely breaks down for flesh-and-blood

'actualized' by concrete beings, but not identical to them. However, this does not answer the question of how things are worse for Tommy in a Tommy-less world.

individuals, such as the previously discussed Tommy, for he would not exist in any sense if he were not created (assuming that such a concrete individual could not have been identical to a merely possible being.)

One possible reply is to say that in order for states of affairs to have value for a being it is enough that the being has a *capacity* for being concrete, an animal, and so on, and both the anencephalic infant and the merely possible being have (some of) these capacities. This reply works at most for merely possible beings; it does not work for the flesh-and-blood Tommy, since he does not exist at all in the Tommy-less world (assuming, again, that he could not be identical to a merely possible being), and thus cannot exemplify any capacities in this world). For this reply to succeed one must assume that one's being such that one possibly has *F* is sufficient for one's having a capacity to have *F*. But this assumption about the relation between modal properties and capacities is far from obvious. For example, from the mere fact that you possibly jump to the moon, i.e., that there is a possible world in which you do this, it does not follow that you have the capacity to jump to the moon.[16] To have a capacity to do or have something (in the ordinary sense of 'capacity') requires more than just having the purely modal feature of being such that one possibly does or has it. Arguably, the capacity must be somehow grounded in features that are not purely modal.

Another problem with the analogy is that it assumes that the lack of conscious experience is worse for the anencephalic than having a conscious good life. However, whether this is true depends on which theory of our identity is correct. To see this, suppose that doctors somehow manage to repair and supplement the anencephalic infant's brain so that it has all the relevant consciousness-generating parts. Assume that our identity is essentially tied to certain consciousness-generating parts of the brain so that an anencephalic infant, which lacks these parts, cannot be identical to a being that has them.[17] Then a *new being*, one of "us", comes into existence when the brain is repaired and supplanted with more brain matter, and things are good for this being. Call the anencephalic infant A and the new being B.

Now, when B begins to exist, either A ceases to exist, or A continues to exist. If A ceases to exist, then A is not better off after the brain reconstruction, since A does not have any valuable experiences after the reconstruction. If A continues to exist, then since A and B are not identical, it is not clear how the fact that B is well-off can explain that A is better off. One option is to say that A is better off in virtue of having B, who is well-off, as a part. More exactly, the idea is that A is conscious in a *derivative* sense—i.e. in the sense of having the essentially conscious B as a part—and this

[16] Johansson (2010) makes this important observation.
[17] See, e.g. e.g., Unger (1990), Unger (2000).

explains why B's pleasant experiences make A better off.[18] The details of this view must be worked out, but it is clearly a contestable view. It is not obvious that something's having a conscious being for whom things can be good (or bad, or neutral) as a proper part is sufficient for *its* being such that things can be good (or bad, or neutral) for *it*.[19]

To see the problem of having too many beneficiaries even clearer, suppose that a human animal is cloned from some of A's cells, and that by altering certain genes, the clone develops without a brain. Call the clone A*. Next, suppose that B is operated on by a procedure that involves the removal and transplantation of the brain from A's skull to A*'s skull. As a result of the operation, A* acquires B as a conscious proper part, and A becomes brainless. Suppose that B benefits from this procedure by experiencing a slight increase in quality of life as a result of the transplant, and that B goes on to enjoy a life that is very good in absolute terms. If a human animal can benefit from acquiring a conscious being as a proper part, then the transfer of the brain from A to A* should render A* *better off*. Similarly, A should be *worse off* as a result of losing B as a part. But it is much more plausible that B is *the only* being in this case who is affected for better or worse.[20]

If B is the only beneficiary in this case, then A* does not benefit by gaining B as a proper part. In that case we should also think that A, the anencephalic infant in Persson's example, would not benefit from gaining B as a proper part. We should instead think that, in this case, B is the only being for whom things have positive value. From this it follows that the anencephalic infant is not better off after the brain reconstruction.

IV

So far, we have scrutinized Persson's argument for the claim that non-existence is worse for a merely possible being than an existence in which things are overall

[18] For defenses of the view that a being can be conscious derivatively in virtue of having a conscious being as a part, see Persson (1999). Persson does not ultimately endorse this view, but defends it against certain objections. Proponents of this view include McMahan (2002: ch. 1), Campbell & McMahan (2016).

[19] Hud Hudson defends a similar claim regarding moral status: having a being with moral status as a proper part is not sufficient for having moral status. See Hudson (2001: 155).

[20] Notice that if A* benefits from the transplant, this cannot be wholly explained in terms of B's benefiting from the transplant and A*'s having B as a proper part. For A*'s benefit and B's benefit could be of different sizes. For B, the benefit might only be very slight, but for A*, the benefit would be enormous—it would be the difference between being permanently unconscious and being conscious with a very good life. But the claim that there are benefits of different sizes in this case is bizarre. It is much more natural to say that the only benefit is the one that is enjoyed by B.

intrinsically good for that being. It seems to us that Persson has failed to establish the conclusion of his argument.

Is this bad news? It would not be bad news if our obligations of beneficence to future generations in non-identity cases can be explained by appealing to their existence being *good* or *bad* for them. For example, we can have a reason to create a person if her life would be good for her. We don't need to say that her existence would be *better* for her than her non-existence.

However, Persson is not convinced that appealing to what would be good or bad for a person (as opposed to what would be better or worse for her) can justify any choice to create or refrain from creating a person. He provides both a specific argument and a general argument. His specific argument is this. Suppose that you can either create A with a good life or give a benefit to a different person, B, who would exist independently of your actions.

> Then if you bring the being (A) into existence, you would have done what most benefits the beings who morally count. But, on the other hand, if you do not bring this being into existence, you would also have done what most benefits the beings (B) who morally count, since now the being (A) that you could have brought into existence does not count morally because, being forever non-existent, it cannot be said to have been harmed by being denied existence. Therefore, you do not have a reason to bring a being into existence rather than not to bring it into existence if it can be benefited only in a non-comparative sense because if you do not act on the reason, it dissolves, and there is no reason to which you have acted contrary. (2017: 59)

This argument is not convincing. Persson assumes that the defenders of a reason to create the non-comparative benefit of a good life must be *moral actualists*. He assumes that they must say that you have a reason to bring about a non-comparative benefit only if the beneficiary actually exists (or will exist).[21] It is true that this view will lead to the problems Persson presents. But one can formulate the relevant reason in non-actualist ways. For example, one can say that if it holds that one can perform an action that would bring about an outcome in which someone will lead a good life, then this fact provides a reason to perform the act. This view will not lead to the problems Persson presents, since on this view whether one has the reason does not depend on who does exist and will exist. It is enough that the relevant

[21] Throughout this section, when we refer to reasons, we have in mind specifically non-instrumental reasons of beneficence—those grounded in considerations of what would benefit individuals. According to Persson, such reasons make up one important class of moral reasons; the others concern autonomy and egalitarian justice. For further details, see Persson (2017: ch. 1).

counterfactuals are true. In the case Persson depicts we can use this account to decide what to do. We know that one action is such that if it were performed, then someone would lead a good life, and if the alternative action were performed, then a distinct individual would exist and lead an even better life. On the basis of these facts, one can say that one has more reason to perform the second act than the first.

Persson's general argument for there being no reason to benefit a being in a non-comparative sense unless one thereby also benefits her in a comparative sense is as follows:

> To ascertain that an action would provide someone with a non-comparative benefit is (...) not sufficient to show that—as far as this individual is concerned—there is reason to perform the action. For it may be that the outcome of performing the action is not better, all things considered for the individual, than the outcome of not performing it because bestowing this benefit removes or prevents the individual's having or getting an *even greater* benefit. Thus, in order to determine that—as far as this being is concerned—you have reason to perform the action, you need to ascertain (...) that the action benefits the being in the comparative sense. (2017: 58, italics added)

We agree with Persson that there is no contrastive reason to give a person a non-comparative benefit rather than do what results in her having *an even greater benefit*. However, this does not show that there is no contrastive reason to give her a certain non-comparative benefit – a pleasure, say – rather than do what results in her not having this benefit. Thus, we think, Persson has failed to establish that the only reason to create a being with a good life is that existence with a good life is *better for that being* than non-existence. We can have a contrastive reason to create a person who will have a good life rather than do what results in her not existing at all, and thus not having any non-comparative benefit.

V

We have scrutinized Persson's argument for the claim that non-existence is worse for a merely possible being than an existence in which things are overall good for that being and we have found it wanting. Far from being "incontestable" the argument's premises, as initially stated, are clearly false, and the argument is logically invalid. Next, we presented what we take to be the most promising reconstruction of Persson's argument. Although the reconstructed argument is valid, the premises are still *highly contestable*. They assume that there are merely possible beings and that things can be better or worse for them. Persson's argument

for the existence claim is not convincing and his argument for the value claim is based on a shaky analogy with anencephalic infants. Finally, we showed that Persson is not justified in his dismissal of the idea that we can explain our obligations to future generations by appealing to facts about their existence being good or bad for them.

In sum, Persson has not made a convincing case for giving merely possible persons membership in the moral club. They will have to look for a different guarantor.

References

Arrhenius, Gustaf & Wlodek Rabinowicz (2010) "Better to Be than not to Be?", in: Joas, Hans & Barbro Klein (eds.) *The Benefit of Broad Horizons: Intellectual and Institutional Preconditions for a Global Social Science*, Leiden: Brill.

Arrhenius, Gustaf & Wlodek Rabinowicz (2014) "The Value of Existence" in: Hirose, Iwao & Jonas Olson (eds.) *The Oxford Handbook of Value Theory,* Oxford University Press.

Balashov, Yuri (1999) "Non-Zero Physical Quantities", *Synthese* 119: 253–286.

Bradley, Ben (2009) *Well-Being and Death*, Oxford University Press.

Bykvist, Krister (2007) "The benefits of coming into existence", *Philosophical Studies,* vol. 135, no. 3: pp. 335–362.

Bykvist, Krister (2015) "Being and Well-Being", in: Hirose, Iwao & Andrew Reisner (eds.), *Weighing and Reasoning*, Oxford University Press.

Campbell, Tim & Jeff McMahan (2016) "Animalism and the Varieties of Conjoined Twinning", in: Blatti, Stephan & Paul Snowdon (eds.), *Animalism: New Essays on Persons, Animals, & Identity*, Oxford University Press, 2016.

Holtug, Nils (2001) "On the Value of Coming into Existence", *The Journal of Ethics* 5.

Holtug, Nils (2010) *Persons, Interests, and Justice*, Oxford University Press.

Hudson, Hud (2001) *A Materialist Metaphysics of the Human Person*, Cornell University Press.

Johansson, Jens (2010) "Being and Betterness", *Utilitas* 22: 285–302.

Luper, Steven (2007) "Mortal Harm", *Philosophical Quarterly* 57: 239–51.

McMahan, Jeff (2002) *The Ethics of Killing: Problems at the Margins of Life*, Oxford University Press.

Persson, Ingmar (1999) "Our Identity and the Separability of Persons and Organisms", *Dialogue* 38: 519–34.

Persson, Ingmar (2017) *Inclusive Ethics: Extending Beneficence and Egalitarian Justice*, Oxford University Press.

Roberts, Melinda (2003) "Can It Ever Be Better Never to Have Existed at All? Person-Based Consequentialism and a New Repugnant Conclusion," *Journal of Applied Philosophy* 20: 153–85.

Stalnaker, Robert (2012) *Mere Possibilities*, Princeton University Press.

Unger, Peter (1990) *Identity, Consciousness, and Value*, Oxford University Press.

Unger, Peter (2000) "The Survival of the Sentient", *Philosophical Perspectives* 14: 325–48.

Williamson, Timothy (2013) *Modal Logic as Metaphysics*, Oxford University Press.

Göran Duus-Otterström[1]

Liability for Emissions without Laws or Political Institutions

Some theorists have recently argued that liability for greenhouse gas emissions presupposes positive law regulating emissions at the time of emitting. According to one account, this is because rights and duties in relation to environmental pollution do not exist before positive law. According to another account, it is because over-emitting actors could not reasonably have known that they were over-emitting until an institution emerged that regulated emissions. The paper claims that these accounts are mistaken. Drawing on the idea that actors had a duty to promote the emergence of just legal regulation, it argues that an actor is typically liable for pre-legal emissions if emitting less would have made just legal regulation more likely, the actor was aware or should have been aware of this, and emitting less would not have been unreasonably burdensome.

[1] Institute for Futures Studies & Department of Political Science, Aarhus University, goran.duus-otterstrom@iffs.se.

1. Introduction

The central question of climate justice is how the burden of averting (more) dangerous climate change should be distributed. An influential answer holds that the burden should be distributed among those who, in over-emitting greenhouse gases, caused the problem. This is the message of the so-called Polluter Pays Principle. While the Polluter Pays Principle cannot supply the whole answer to just climatic burden sharing—it does not apply to the emissions of dead polluters, for example—it is widely thought that living, culpable and affluent actors should pay for having polluted, at least when the emissions exceeded the actors' fair share of the atmosphere's absorptive capacity.[2]

The pollution-centered view has recently been called into question by Carmen Pavel (2016) and Paul Bou-Habib (2019), who in a pair of interesting papers argue that the idea of making polluters pay for past emissions overlooks the role of positive law for liability. More specifically, they argue that it is largely impermissible to hold actors morally liable for emitting greenhouse gases unless the emissions were in breach of legally promulgated duties. For Pavel, this is because pre-legal liability would apply a nonexistent normative standard. As she puts it, 'there are no moral entitlements with respect to pollution prior to legal conventions that establish them' (Pavel 2016, 337). For Bou-Habib, by contrast, the point is that climate change is so complex that actors cannot reasonably figure out that they are over-emitting on their own. Institutions that authoritatively set out legal rights and duties are needed fairly to hold actors liable for emitting. The shared conclusion is that we should place little weight on historical emissions in deciding how the costs of climate policy should be shared since it is not until relatively recently that there has been an international legal regime for regulating GHGs (if there is one even now).

These are important and novel arguments. They complete and crystallize, in different ways, the skepticism that has been brewing in the literature about the idea of holding actors liable for emissions in the absence of political institutions.[3] It should immediately be pointed out that neither Pavel nor Bou-Habib rely on the (implausibly strong) view that liability always presupposes positive law. They do not

[2] Methodologically, this debate treats the climate challenge as a question of distributing remedial responsibility (Miller 2007, 98). It is not obvious that this is the right approach, but I will not question it here. For general discussion of the problems of Polluter Pays Principle (PPP), see Roser and Meyer (2010) and Gardiner (2011, 414–20). For PPP as a fault-based principle, see Shue (1993) and Vanderheiden (2008). For PPP as a forward-looking principle, see de Sadeleer (2002). For PPP and excusable ignorance, see Vanderheiden (2008 ch. 6) and Bell (2011a). For PPP and dead emitters, see Caney (2005) and Duus-Otterström (2014). A popular idea is to supplement PPP with the Ability to Pay principle; see Caney (2010) and Page (2011). Bowman (2019) questions the proportionality assumption that gives rise to the need to supplement PPP.

[3] See, e.g., Miller (2008), Meyer & Sanklecha (2014), Blomfield (2016).

take issue with, for example, holding people liable for having owned slaves even though owning slaves was not illegal at the time.[4] But they think, not implausibly, that climate change is importantly different from a case like slavery since emitting greenhouse gases is neither invariably wrong nor a clear violation by one actor of another actor's moral rights. The causes of climate change are more amorphous and more morally benign, and this makes liability for pre-legal actions more problematic.[5]

The aim of this paper is to argue that Pavel and Bou-Habib's challenge fails. The response I develop draws on John Rawls's idea that there is a natural duty to promote the emergence of just institutions. The argument proceeds in two steps. I first argue that if a state in the pre-legal situation failed to promote, or helped prevent, the occurrence of international laws regulating emissions, then the state can rightly be held liable for this.[6] I then argue that since states can fail to promote such laws by not curbing their emissions, we can rightly take pre-legal emissions into account in discussions of burden sharing. I offer this argument—which is intended to be broadly internal to the Pavel and Bou-Habib's projects while avoiding the familiar charge that the natural duty of justice cannot be rendered concrete enough to support ascriptions liability—in section 4. I begin by offering a closer presentation of Pavel and Bou-Habib's challenge.

2. Positive Laws and Liability

At its core, Pavel and Bou-Habib's challenge to the pollution-centered view of climate justice draws on the principle I call:

> *Laws before Liability for Emissions* (LLE). An actor's liability for greenhouse gas emissions presupposes that there was, at the time of emitting, legitimate positive law determining the amount of emissions that the actor was entitled to emit.

I will look closer at how Pavel and Bou-Habib defend and qualify this principle in sections 3 and 4. In this section, I specify the principle and explain why it is potentially highly important.

Let us begin by explaining the key terms. 'Liability' should here be understood as

[4] Bou-Habib uses the example of slavery (2019, 1303). Pavel uses assault as an example of rights violations in the absence of legal conventions (2016, 341).
[5] Bou-Habib approvingly quotes David Miller's remark that 'Global warming is not like slavery, where there was a clear historic wrong that required, and may still require, redress' (Miller 2009, 136).
[6] Bou-Habib notes that 'political obstructionism' challenges his argument (Bou-Habib 2019, 1308), but I argue that he underestimates the extent to which it does so. I return to this issue in section 4.

the responsibility to take on a cost or burden: an actor is 'liable' for causing an outcome insofar as the actor has a responsibility to take on some cost or burden in virtue of having caused the outcome, where 'responsibility' is used in a normative sense to indicate that the actor either has no moral right to reject paying or, more strongly, should pay.[7] Importantly, liability thus understood is not the same as actually paying for having caused the outcome. We can be liable without being held liable and vice versa. Liability is also not a legal concept. The point of LLE is that actors are not *morally* liable for emissions until there are legitimate laws governing emissions, meaning that actors are not cost-responsible for these emissions. It can be tricky to get one's head around this point since holding actors liable is often the same thing as holding them legally liable, that is, imposing on them a legal responsibility to take on a cost or burden. But the simple idea behind LLE is that we must not hold actors legally liable unless they are morally liable. It thus rules out, among other things, holding actors retroactively legally liable for emissions that occurred before emissions were legally regulated.

LLE expresses a general relationship between emissions liability and positive law and is just as applicable to the domestic level as it is to the global level.[8] But discussions about climatic burden sharing are typically aimed at the international level, and what the principle then rules out is specifically the idea of holding states liable for emissions that occurred before there was an international treaty setting out states' entitlements to emit in the form of legally binding mitigation commitments or emissions quotas.[9] This idea stands in stark opposition to proposals based on liability for pre-treaty emissions, such as the famous Brazilian Proposal. The Brazilian Proposal, which Brazil injected into the negotiations leading up to the Kyoto Protocol, held that states should be assigned climate mitigation targets as a function of their cumulative emissions since 1850, the argument being that states that had contributed more to climate change should be given greater remedial burdens. LLE deems any such approach misguided. For most of human history, the

[7] Liability can be fault-based or strict depending on whether causation is sufficient for liability (Vanderheiden 2008). The sense of liability invoked by LLE is fault-based, but it need not draw on thicker moral notions such as moral blame (Shue 1999; cf. Miller 2007, 86–90). For a paper discussing retroactive liability for emissions from a legal perspective, see Farber (2017).

[8] I understand laws in a positivist way throughout the paper. Taking a natural-law approach would challenge LLE in ways not addressed here (Marmor 2011).

[9] Since we are now speaking of international treaty law, it is clear that 'legally binding' does not mean 'enforceable' or 'enforced;' it only means that the treaty sets out mandatory emissions quotas to states. But it is worth noting that there are dimensions that need further development once LLE is applied to the international level. First, what is the relevance of having approved the treaty? For example, if 'only' 95 percent of the world's states were to ratify an international treaty, would liability be limited to those states or would the remaining 5 percent also be liable for their emissions? Second, what is the relevance of domestic laws for international liability? If a state is overshooting unilaterally adopted emissions target, would the state be liable to other states? These are important questions, but since they do not pertain to the core claim of LLE I set them to one side.

atmosphere was an unregulated commons which could be freely used, and since liability presupposes legitimate law *at the time of emitting*, it would be impermissible to hold states liable for emissions they undertook during that period.

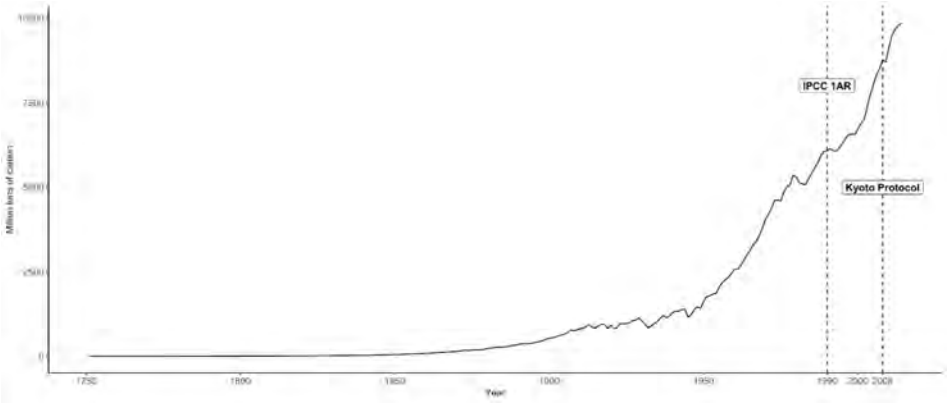
The reference to 'legitimate' law is crucial because it indicates that, for proponents of LLE, breaching a legally promulgated emissions duty is not sufficient for being liable for emitting. A legally binding emissions treaty only grounds liability for emissions if it meets a standard of legitimacy. For Pavel this means that the treaty must have been adopted through a reasonably impartial, inclusive, and not seriously epistemically defective political process (Pavel 2016, 361). Bou-Habib (2019, 26), relying on Buchanan and Keohane's (2006) well-known account of the legitimacy of global institutions, argues that the treaty must amount to, or flow from, institutions that are minimally morally acceptable, beneficial compared to feasible alternatives, and maintain integrity.

LLE has significant ramifications for the way climate policy costs ought to be shared. A widely endorsed view in the climate justice literature is that actors are not liable for emissions undertaken in excusable ignorance of anthropogenic climate change. This is often taken to mean that states are not liable for emissions prior to 1990, the year of the Intergovernmental Panel on Climate Change's first Assessment Report. LLE, however, threatens to remove a large chunk of post-1990 emissions as well. Bou-Habib thinks that the first legitimate and legally binding climate treaty was the Kyoto Protocol and consequently maintains that liability for emissions started in 2008. That would exempt states from liability for up to 119,500 million metric tons of carbon compared to the standard approach of holding states liable for emissions from 1990 and onwards (Figure 1).

Pavel for her part questions the legitimacy of the Kyoto Protocol and notes that the preconditions for liability might *still* be missing at the international level. That would of course have even more significant implications than Bou-Habib's more optimistic assessment. Both Pavel and Bou-Habib reject that the preconditions of emissions liability began with the 1992 United Nations Framework Convention on Climate Change on the grounds that the Convention did not include legally binding emissions quotas.[10]

[10] Bou-Habib (2019, 1307) makes this point explicitly; Pavel (2016, 362) makes it implicitly.

Figure 1: Global Carbon Emissions from Fossil-Fuel Burning, Cement Manufacture, and Gas Flaring, 1751–2014 (million tons of carbon)



Comment: Data from Boden et. al. (2017). 1990 is often considered the year when ignorance of anthropogenic climate change was no longer excusable. 2008 marks the beginning of the first commitment period of the Kyoto Protocol.

But LLE undermines not only pollution-based approaches to climatic burden sharing. In exonerating actors of many past emissions, it also undermines benefit-based ones (Blomfield 2016; Bou-Habib 2019). Some climate ethicists argue that current people are cost-responsible for previous generations' emissions insofar as they possess the fruits of the emissions (Shue 1999; Roser & Meyer 2010; Page 2012; Goodin 2013; Duus-Otterström 2014). But if previous generations did nothing wrong in emitting as much as they did (because the emissions were not legally regulated at the time) then it is more difficult to see why current actors would be cost-responsible for possessing the benefits that their emissions created.

3. Laws before Liability for Emissions: Two Defenses

LLE asserts that moral liability for emissions presupposes that emissions were legally regulated at the time of emitting. An actor is not liable for emitting unless the emissions breached a duty laid down by legitimate law. In this section, I consider Pavel and Bou-Habib's arguments for this principle. In section 4, I explain why these arguments are unsuccessful even if we accept their contention that knowing and avoidable contribution to harmful climate change is insufficient to ground liability.

3.1 Pavel's conventionalist defense

Pavel's argument is one of *legal conventionalism*. Legal conventionalism is the view that 'duties and rights ... arise out of legal conventions, that is legal decisions made by the community or the relevantly situated people in that community (judges, legislators)' (Pavel 2016, 343). Pavel accepts, however, that 'we have rights and obligations to treat each other in certain ways by virtue of our common humanity that both precede and transcend political communities' (ibid., 343). What she rejects is that environmental pollution is a case in which these pre-political rights and obligations apply. Her view is thus conventionalist *with respect to pollution*. She thinks that 'we cannot hold people responsible for polluting without a system of legal rights in place that assigns entitlements, protections, and obligations' (ibid., 338).

The key notions in Pavel's argument are the ideas of balancing legitimate interests and multiple equilibria. Pavel assumes that we are only liable for inflicting harm on others if the harm amounts to a rights violation. But she argues that to determine whether a harmful activity amounts to a rights violation in a case like environmental pollution, it is not enough to look only at the harm caused by the activity; we must also look at the benefits people gain from engaging in it. People have a legitimate interest in 'driving cars, running factories or producing energy' (ibid., 344). Hence it would be excessive to deem all pollution as impermissible just because it is harmful. What we instead need to do is balance the interests we all have in polluting and in avoiding environmental harm.

Pavel thinks that the rights relating to pollution arise from the balancing of legitimate interests. Thus, for her, the reason legal conventionalism is appropriate with respect to pollution is simply that pollution cannot amount to a rights violation until the balancing has occurred. This makes pollution importantly different from cases in which rights are violated irrespective of legal conventions, such as assault. Those who assault others are liable because they violate pre-political moral rights. Pollution, by contrast, belongs to a class of harms comprising 'tragedy of the commons problems and other scenarios in which people have an interest in engaging in activities that cumulatively have a tendency to cause harm, but prohibiting these activities causes harm to the people who have an interest in engaging in them' (ibid., 342). Here legal conventions are needed to '*create and specify* rights and obligations with respect to pollution' (ibid., 343, italics in original), and this is so even though pollution might set back interests protected by our moral rights, such as bodily integrity or property rights. The fact that pollution sets back interests protected by moral rights is not enough to show that it is a rights violation since, again, we also have a legitimate interest in engaging in activities that pollute.

The idea of balancing is crucial for Pavel, because if harmful pollution were always morally impermissible, it is not clear why legal conventions would be needed for liability. Yet balancing as such is not enough to establish legal conventionalism since we might just think that the line between permissible and impermissible emissions can be drawn without legal conventions, for example, by scientists or philosophers. Pavel argues, however, that legal conventions are necessary since there are many ways in which the balance between polluting and harm could be struck, none of them more correct than the others, and since the 'kind of judgments for fine-tuning different trade-offs required by considering the relative importance of different interests and harms in the case of pollution cannot be the result of abstract moral theorizing' (ibid., 357). Pavel's argument, then, is that since the issue concerns striking a balance between actors' interests, and since there are 'multiple equilibria' in doing this, how the balance is struck must be up to the actors, via some legitimate lawmaking process.

What follows from this is a strong limitation on relying on a principle like the Polluter Pays Principle in the context of climate change. The Polluter Pays Principle holds that actors should be held cost responsible for emitting too many greenhouse gases. Pavel's response is that this position is incoherent until there are legal conventions regulating the extent of actors' emissions. Pollution is only grounds for liability if it violates rights, and our pollution rights are created and specified by positive law. Thus, until there are legal conventions, the notion of 'too many greenhouse gases' is vacuous.

I give my response to Pavel in the next section. Here I just want to flag some general questions her argument faces. First, while Pavel repeatedly writes that we 'cannot' hold actors liable for pre-legal pollution, it is not clear whether she means this literally or if it is just another way of saying that doing so would be inappropriate or incoherent. It is most reasonable, however, to make the latter reading. Thus understood the point is not to deny that we could hold someone cost responsible in the absence of law but rather that doing so would in an important sense be premature: since laws are constitutive of rights and obligations regarding pollution, we simply lack the conceptual tools to say that an actor's pre-legal emissions are wrong. However, second, while Pavel seems to understand her view as ruling out retroactive application of laws, this conclusion does not follow from her premises. Pavel writes in the present tense: she says that polluters are 'only responsible if they *produce* effects above a threshold defined in the law' (ibid., 354. Italics added.). But we could accept the claim that environmental rights and obligations depend on legal conventions while thinking that the conventions could be applied retroactively *once*

they emerge.[11] To make her argument speak against retroactivity, then, Pavel would have to add a component explaining why retroactive application of legal conventions would be wrong. The obvious choice would be to stress the importance of giving actors a fair warning. Third, Pavel's argument seems to assume that the interests in harming and not being harmed by pollution are symmetrically distributed. Her argument is at least most compelling when pollution benefits and harms everyone roughly the same. Yet climate change might be a case where the harms mainly befall people who themselves benefit little from the pollution. This kind of asymmetry can cause problems for non-consequentialist forms of justification (Ashford 2003, 294–301).

3.2 Bou-Habib's epistemic defense

Unlike Pavel, Bou-Habib (2019) accepts that there is an answer as to what each actor was morally entitled to emit before the emergence of legitimate laws governing emissions. Thus, for him, the problem with pre-legal liability is not that it holds actors to a nonexistent standard of excessive emissions. The problem is rather that liability would be unfair in the absence of certain epistemic conditions. Bou-Habib believes that 'past actors who could not reasonably have been expected to know that they were emitting excessively did not acquire original duties of compensation merely on account of their having undertaken excessive emissions' (ibid., 1305). Thus, in addition to excusable ignorance of the phenomenon of anthropogenic climate change, he posits that there is excusable ignorance of one's own *entitlements to emit*. His argument is (i) that actors should not be held liable for excessive emissions if they could not reasonably have known that their emissions were excessive, and (ii) that actors often could not reasonably have known that their emissions were excessive until legitimate institutions emerged. The fact that Bou-Habib speaks of 'institutions' rather that 'laws' need not detain us because the role of institutions in his account is mainly to promulgate legally binding duties to states.

While not universally accepted, many agree that it would be unfair to hold actors liable for emitting when they were excusably ignorant of climate change, particularly since doing so would impose burdens on actors who did not enjoy a fair opportunity to avoid liability.[12] Bou-Habib is right to point out that the same concerns might speak against holding actors that were excusably ignorant of their own over-emitting liable. The more controversial part of his argument, then, is the claim that

[11] I am grateful to Lukas Meyer for pointing out that Pavel's argument is compatible with retroactivity.
[12] Hart (2008) offers a classic defense of giving actors a fair warning in the context of legal punishment. But see Caney (2010), Bell (2011a) and Gardiner (2011) for a more nuanced (and critical) view of the role of excusable ignorance in climatic burden sharing.

legitimate institutions are necessary to remove excusable ignorance of one's own over-emitting. The pivotal step in the argument here is to do with the ability of institutions to reduce 'social complexity,' by which Bou-Habib means 'an evolving condition of a society in which its members affect each other's outcomes in increasingly significant ways along multiple and interconnected causal pathways that are difficult to discern' (ibid., 1300–1301). Bou-Habib thinks that given the great social complexity of climate change, it would be unreasonable to expect states to figure out their entitlements to emit greenhouse gases on their own. International institutions that authoritatively specify the entitlements to emit are needed. It follows that, absent institutions, we cannot fairly hold states liable for their emissions. Even states who were aware of anthropogenic climate change should be exonerated for emitting, provided that their emissions were not extreme enough to fall outside the 'gray area' of reasonable disagreement (ibid., 1306).[13]

Like Pavel, Bou-Habib notes that institutions are not necessary for liability across the board. For example, it would not be unfair to hold slave owners or murderers liable for having enslaved or murdered people during a time when such acts were legally prohibited. But murder and slavery are clear *mala in se* in which social complexity is exceedingly low. Reasonable actors did not need institutions to know that slavery or murder was unjust—everyone should have known that owning one slave was already one too many. The case of climate change is different, Bou-Habib argues. Here social complexity is so high that pre-institutional liability would be unfair: even though an actor in the pre-institutional setting may *in fact* be emitting too much, it could not reasonably be expected to *know* this given that there is reasonable disagreement over both the overall aims of climate policy (e.g., the global temperature target) and the right way of allocating the burdens of meeting that aim (ibid., 1306).

There is a tension in Bou-Habib's argument that should be registered upfront. His argument assumes that there is a correct answer to whether someone emits excessively; institutions do not *create* the very notion of 'excessive emissions' like on Pavel's account. This naturally raises the question of whether legitimate institutions are also *sufficient* for liability. Consider this passage:

> 'the conclusion that historical climate duties arise from past excessive emissions is only secure for all past excessive emissions from the moment after which legitimate institutions of global climate governance promulgated emissions

[13] Bou-Habib gives this negative argument to undermine what he calls the 'preinstitutional liability claim' (2019, 1303). His positive argument is that the absence of legitimate institutions does challenge the legitimacy of the economic status quo (cf. Caney 2006; Blomfield 2016). I set Bou-Habib's positive argument to one side since I am interested in challenging specifically his negative argument.

> duties to states, since it is only after this moment that we can confidently say of past actors that they should have known that they were breaching their emissions duties' (ibid., 1304)

The reference to 'emissions duties' is unclear. Are these the moral duties states have pre-institutionally or the legal duties that institutions create? Given that Bou-Habib wants to offer an epistemic argument, the answer should presumably be the former. But then we immediately see the tension: why assume that the institutions track states' pre-institutional moral duties as opposed to promulgating legal duties that are out of whack with them?

There are two ways Bou-Habib could respond. He could argue that legal duties generate moral duties. On this view, regardless of whether an institution tracks an actor's pre-institutional moral duties, once the institution generates legal duties, the actor has a moral duty to discharge these duties. This would be like saying that actors incur content-independent political obligations to reduce emissions once a legitimate institution requires it.[14] Or he could argue that there is a determinate list of potentially just institutional schemes even though actors have no way of telling which of these schemes is correct.[15] In such circumstances one could argue that, for any scheme selected, the actors must take it as specifying their moral emissions duties since each scheme specifies pre-institutional moral duties just as plausibly as any other scheme on the list.

Bou-Habib does not explain how he proposes to avoid the tension between moral and legal duties. My sense, however, is that he assumes something like the first response. That would be fine, but it is worth noting that it would weaken the link to epistemic considerations, for the role of institutions is then restricted to coordinating the efforts of actors in reasonable disagreement. Opting for the second, semi-procedural option would allow a clearer link to epistemic considerations. In any case, since it would be implausible to assume that institutions somehow magically manage to track the pre-institutional moral truth, something needs to be added in order to complete Bou-Habib's argument for LLE.

4. The Promotion Argument

It is worth pausing to take in the controversial implications of these arguments. The traditional way to challenge liability for past emissions is, again, to stress excusable ignorance of climate change. Even though states or their populations were contri-

[14] For a good introduction to the vast literature on political obligation, see Horton (2010).
[15] This move resembles Klosko's (2004) semi-procedural theory of political obligation, which in turn is inspired by Rawls (2005), especially Rawls's notion of the burdens of judgment.

buting to dangerous climate change in the past, so this argument goes, for a long time they could not reasonably know this, and so it would be unfair to hold them liable for doing so (Vanderheiden 2008; Bell 2011a; Farber 2017). LLE makes the stronger claim that even after states or their populations became aware that they were contributing to dangerous climate change, there is still no basis for holding them liable for emissions until laws regulating emissions are adopted. If we make the plausible assumptions that it was within the power of states to emit less and that emitting less would not have been overdemanding, this is tantamount to saying that it would be unfair to hold actors liable for knowingly causing reasonably avoidable harm to others. We might think that this simply cannot be unfair.[16]

LLE also has the counterintuitive implication that liability for emissions could go away again. Suppose the Paris Agreement counts as a legally binding and legitimate treaty. Suppose further that the Paris Agreement were to implode, pushing international climate politics into a state of anarchy until a new treaty comes into force 2030. LLE then entails that any emissions between now and 2030 could not rightly serve as a basis for liability. But it can be difficult to see why states would not be liable for emissions emitted in the coming decade just because there is no longer a legally binding and legitimate climate treaty. The emissions must be judged against the backdrop of an impending and scientifically established climate disaster. We might think that the 2030 treaty can and should hold states legally liable for what they emitted.

These problems are mitigated by the fact that neither Pavel nor Bou-Habib argue that laws are strictly necessary for liability for emissions. Their arguments admit that we could fairly hold actors liable for *clearly* excessive pre-legal emissions. Bou-Habib is explicit about this. He notes that those who undertook 'extremely high emissions' are liable because they can 'be expected to have known that they were breaching their emissions duties' (2019, 1303). But Pavel's view also leaves some space for pre-legal liability since some ways of striking the balance between pollution and harm are simply 'beyond the pale' (Pavel 2016, 361). There is no reason to think, for example, that pollution that completely or almost completely discounts harm must await social balancing before we can reject it as morally wrong. Pavel wants to handle such cases via her criteria for legitimacy, which among other things call for inclusive legislative procedures (Pavel 2016, 361). But we are arguably able to identify some levels of emission as *substantively* excessive independently of procedures. For example, it seems substantively excessive to emit at a rate twenty times greater than what would be consistent with avoiding dangerous climate

[16] To be precise, such liability can be *comparatively* unfair but it does not seem *noncomparatively* unfair (Feinberg 1974). For an argument that it would be wrong to cause (expected) harm by emitting reasonably avoidable emissions, see Hiller (2011).

change if emitted by everybody. Pavel's legal conventionalism, then, does not rule out that *some* emissions profiles are grounds for pre-legal liability.

Since extreme emissions call for a different conclusion, it is more accurate to say that Pavel and Bou-Habib are drawing on:

> *Laws before Liability for Non-Extreme Emissions* (LLE*). An actor's liability for non-extreme emissions presupposes that there was, at the time of emitting, legitimate positive law determining the amount of emissions that the actor was entitled to emit.

When we restrict the discussion to non-extreme emissions, it may seem that Pavel and Bou-Habib have a convincing case after all. Few people doubt that *some* level of emission is morally permissible; the relevant consideration when it comes to liability for emissions is whether someone has over-emitted, not merely whether they have contributed to harm. But how could we tell that a non-extreme emitter is 'over-emitting' unless laws or political institutions regulating the atmosphere are in place? The alternative seems to be that states should anticipate—and be held liable for exceeding—the quotas they *would* enjoy under just a climate treaty. But that idea will look dubious to anyone who feels the force of Pavel and Bou-Habib's arguments. Why assume that there is a pre-legal answer as to how emissions should be allocated, or expect states to have epistemic access to it?[17]

Part of the discussion will of course be determined by the share of emissions qualifying as 'extreme.' LLE* will not speak against the idea of allocating climate policy costs based on past emissions if most past emissions were in fact extreme. But Pavel and Bou-Habib reject that most emissions were extreme; they attack the Polluter Pays Principle precisely because they think that most emissions fell within a range where laws were required to say how many emissions were too many. To convince people who agree with this, just stressing the duty to avoid contributing to harm will not do. Instead, we must explain why states can be liable for pre-legal emissions even if (1) the emissions are non-extreme and there either (2a) is no truth to the matter as to what a just level of emissions would be or (2b) states cannot be expected to know what this level is. I offer such a response in what follows. The response only assumes that actors have a duty to promote, or at least not obstruct, the *emergence* of just legal regulation. Positing such a modest duty, I argue, is enough to reject LLE (and LLE*).

[17] This is the difficulty with Bell's otherwise insightful discussion of promotion duties and climate change. Bell suggests that actors in a pre-legal situation are morally required to 'reduce their greenhouse gas emissions to a level that they can reasonably believe would be consistent with the specification and allocation of duties by effective institutions' (Bell 2011b, 115).

4.1 Liability and promotion duties

The fundamental problem with LLE (and LLE*) is that it overlooks that there is a moral obligation to *set up* legal regulation such that actors may be liable for failing to do so. This idea could be explained in several ways, but it is often—and in my view most powerfully—subsumed under John Rawls' account of the natural duty of justice. In what follows, I assume that Rawls offers a plausible way of speaking about these issues.

Rawls famously posited a natural duty of justice to 'to support and to comply with just institutions that exist and apply to us' (Rawls 1999, 99). The duty to support and comply with just institutions is a natural duty, Rawls maintained, because it applies to everyone regardless of voluntary transactions such as consent or promises. Just like everyone has a duty not to be cruel, everyone has a duty not to undermine or disobey just institutions that pertain to us.[18] Rawls recognized that we sometimes find ourselves in situations where just institutions are yet to be established. What the natural duty of justice then requires is that we work towards *establishing* just institutions. As Rawls put it, each person has a duty to 'further just arrangements not yet established' (Rawls 1999, 99). Since this kind of duty is about promoting the emergence of just institutions when there are none, we may refer to it as the 'duty to promote.'[19] Rawls added that our promotion duties are tempered by an overdemandingness proviso. We should not be held liable for failing to promote institutions in ways that would involve 'too much costs to ourselves' (ibid., 99).

Promoting duties allow for a decisive argument against LLE (and LLE*). The first step of the argument is simply to note that if an actor in the pre-legal situation failed to promote, or helped prevent, the occurrence of international laws regulating emissions, then the actor can rightly be held liable for this precisely because it amounted to a breach of the actor's promotion duty. The second step is to realize that an actor can breach their promotion duties by failing to curb emissions. The upshot is that we can rightly take pre-legal emissions into account in current discussions of burden sharing. States may be liable for pre-legal emissions insofar as their emissions helped prevent, or failed to promote, the emergence of laws regulating emissions.

To anticipate an objection, it may seem that the argument falls victim to the

[18] It is worth noting that, for Rawls, natural duties were strictly for individuals (Rawls 1999, 99–100). In what follows, I write as if states or governments have natural duties. Readers who find this awkward can just imagine that states or governments 'have' natural duties because they are populated by individuals that have these duties.

[19] Cripps (2013, 116) uses the same term to denote individuals' duty to enable collective action. This is part of what I mean by 'promoting duties', but on my usage such duties must be directed at creating legal regulation. For a critical discussion of promotion duties in non-ideal circumstances, see Valentini (2017).

standard objection that natural duties of justice are too vague. Promoting duties leave it to the actors' discretion *how* they go about promoting just regulation and thus cannot, it may be thought, sustain anything like liability for emissions, let alone a determinate cost-sharing principle like the Polluter Pays Principle. Yet the argument I offer is intended to show that promoting duties do have quite determinate implications for emissions even when we concede that actors could have discharged those duties in different ways. Indeed, I shall argue that we get something close to a principle allocating climate policy costs according to past emissions from positing promoting duties. The novel part about the argument lies not stressing in promoting duties but in showing how these duties sustain the idea of liability for pre-legal emissions in the context of climate change.

I will not spend much time defending the first step of the argument since I take it to be uncontroversial that actors can have duties to work towards legally regulating that which ought to be legally regulated and may be liable for failing to discharge these duties. But to see why it is plausible that actors can be pre-legally liable for what we might call their 'climate behavior,' it is helpful to consider the case of the United States. After having been authoritatively informed about anthropogenic climate change in 1990—arguably significantly earlier[20]—the United States increased its total emissions of CO_2e by roughly 16 percent 1990-2007 and maintained per capita CO_2 emissions of around 19 tons throughout the period (Boden et al. 2017). In addition, the country badly crippled the Kyoto Protocol by not ratifying it and helped bring about the failure at the Copenhagen climate summit (Gardiner 2011, 127–40; Keohane and Victor 2011). Meanwhile, the George W. Bush administration was seeding doubts about the veracity and urgency of climate change (Vanderheiden 2008, 15–44, 197–206). Given this track record, it would be odd to insist that the United States could not rightly be held liable just because there was no legally binding climate treaty. The United States was instrumental in undercutting the very legal regulation that would have put it in a position to be held liable in the first place. If we believe that the country was under a duty to promote legal regulation of climate change, we must be prepared to take this obstruction as a basis of fair liability.[21]

[20] The Johnson administration received a scientific warning about climate change as early as 1965 (Gardiner 2011, 78). A side note is that it is not clear that United States is liable for its emissions even now according to LLE. The country never ratified the Kyoto Protocol and the current administration has announced its intention to pull out of the Paris Agreement.

[21] I am not suggesting that joining the Kyoto Protocol would have been painless for the United States. Victor notes that the country put itself in a bind by agreeing to an 'unachievable' emissions reduction by seven percent in the run up to the treaty (Victor 2011, 207); Sunstein (2007) argues that the Kyoto Protocol would have been economically detrimental to the United States. But saying that the treaty would have been costly does not automatically show that the treaty would have been unreasonably demanding such that it went beyond the United States' promotion duty.

Yet it is one thing to say that states can be liable for failing to promote legal regulation of climate change. To respond to LLE (and LLE*) we must also show that states can be liable for their pre-legal *emissions*. This second step of the argument is more complicated, but in the abstract the answer is clear. Emissions are a way in which states can fail to discharge their promotion duties. More specifically, a state can make legal regulation of climate change less probable by refusing to curb emissions and more probable by curbing emissions.

How can emissions affect the emergence of legal regulation? Three mechanisms can be highlighted. First, climate change is a paradigmatic collective action problem; it can only be avoided if sufficiently many states contribute to a solution, yet there is an incentive for each state to hang back, partly because of the fear of being a 'sucker.' In such situations, by reducing emissions, a state can assure other states that they will not be put at a relative disadvantage by choosing to cooperate. Second and relatedly, in reducing emissions a state can communicate a readiness to co-operate. Reducing emissions can be seen as a 'costly signal,' that is, a signal that demonstrates an earnest willingness to combat climate change. Third, the state may serve as a role model or template for how a transformation to lower emissions could occur. This is especially relevant for high-income industrialized countries, where the mere act of moving to lower emissions without unreasonable sacrifice is thought important because it shows that a low-carbon future is possible.[22]

Since states' emissions are relevant for the prospects of international legal regulation, they are *directly* relevant to whether states are discharging their promotion duties. This is important because it shows that Bou-Habib's approach to promotion duties is unsatisfactory. Bou-Habib agrees that states had a duty to promote legal regulation of climate change. He writes that 'it is unreasonable that states should escape liability for their excessive emissions if they themselves prevent legitimate institutions from being established and from promulgating emission duties to them' (2019, 1308). But he does not think that this offers a challenge to his argument since 'not all states that emitted excessively after 1990 can be accused of obstructionism' (ibid.). This point automatically goes through if we divorce the degree to which a state emits from the degree to which it obstructs. Once we realize that emissions are directly relevant for promotion duties, however, the picture is more complicated. Some states may be 'obstructing'—preventing the emergence of legal regulation—simply because they are emitting excessively.

[22] The collective action problem at the heart of climate change is laid out well by Barrett (2007) and Victor (2011). The idea that going first takes away some reasons for others to hang back is defended by Shue (2011). For costly signaling in international relations, see, e.g. Gartzke et al. (2017). The idea of leading by being a role model is laid out by Parker & Karlsson (2010) under the heading of 'directional leadership.'

Of course, since we are now deriving liability for emissions from a natural duty to promote legal regulation, we should not look at emissions in isolation. As I argue in the next section, we should rather look at the total package of a state's climate-related behavior. But before we develop the promotion argument further, let us first consider how the argument in its basic form squares with the conventionalist and epistemic defenses of LLE (or LLE*). The conventionalist defense is vulnerable to the argument since stressing promotion duties is not the same thing as preempting the answer as to how emissions should be allocated. The main appeal of the conventionalist defense is that it questions the practice of holding actors liable for emitting 'unjustly many emissions' when the prior question of just allocation has not been settled. Critics are no doubt correct that proponents of pre-legal liability uncritically tend to assume something like an equal per capita right to emit (Blomfield 2016; Bou-Habib 2019). But the promotion argument does not, as such, invoke a full standard of justice in the distribution of emissions. All it says is that states whose emissions fail to take us closer to just legal regulation may be held liable. Hence, it does not fall victim to the critique that we must not hold actors liable according to a non-existent standard.

The epistemic defense is also vulnerable to the promotion argument. The epistemic defense holds that institutions are needed since it is unclear what our entitlements are in a complex issue like climate change. Yet promotion does not pose nearly as severe an epistemic problem as figuring out whether one emits within one's moral entitlements. Here the question is simply whether a state has taken reasonable steps to make the emergence of legal regulation more likely, and while I do not doubt that a detailed answer can be quite complex here too, it is clear that significantly *increasing* one's emissions did not, as states were no doubt aware, help the cause of getting an international climate treaty curbing climate change. This is an important result since increasing emissions was the general trend during the 1990s and much of the 2000s. It would be quite difficult to argue that states cannot fairly be held liable for post-1990 emissions on the grounds that they took every reasonable step to promote just regulation of climate change.

In sum, the promotion argument supplies a robust response to LLE (and LLE*). I do not suggest that it is the only response we could give. For example, as noted above, we could simply argue that actors may be held liable for knowingly inflicting reasonably avoidable harm on others (Hiller 2011). It is fair to say, though, that the promotion argument is closer to the intuitions that sway people like Pavel and Bou-Habib and in that sense is less of an external critique. The argument only requires that we accept the uncontroversial idea that actors have duties to work towards legally regulating what ought to be legally regulated. As long as we bear in mind that the relevant sense of 'over-emitting' draws on whether one has done enough to

promote the emergence of legal regulation of emissions, this gives us what we need to hold states liable for having over-emitted.

5. Discussion and Objections

The general thrust of the promotion argument is that a state is liable for pre-legal emissions if (i) emitting less would have made just legal regulation more likely, (ii) the state was aware or should have been aware of this, and (iii) emitting less would not have been unreasonably burdensome. However, given that the source of liability is a failure to promote, we cannot say that a state is automatically liable for emissions as soon as conditions (i)-(iii) are met. The duty to promote requires taking reasonably demanding steps to bring about legal regulation, but it does not specify exactly what these steps are. It yields a promotion 'quota' which can be met in different ways. The duty to promote therefore admits than a high-emitting state can compensate for emitting by promoting legal regulation in other ways, much like when a climate activist is flying across the world to mobilize people.

Since the promotion argument only requires that actors meet a promotion quota, it may be wondered why past emissions are relevant for burden sharing after all. The quota suggests that states had discretion in the way they went about promoting legal regulation. But the point is that states generally *did not* engage in the sort of behavior that would have justified their failure to significantly curb emissions. It is plausible, therefore, to treat their refusal to curb emissions as a source of liability—and to take their emission records as a rough approximation of the extent to which they have violated their promotion duty. A more detailed analysis would no doubt need to consider the extent to which states violated the duty to promote in ways other than emitting. For example, the promotion argument entails that we should attribute more liability to a state that refused to curb emissions *and* engaged in disinformation about climate change than a state that merely refused to curb emissions. But this point does not change that emission records are relevant for current burden sharing in their own right.

Given that the source of liability is a failure to promote, the promotion argument will not treat all emissions the same. The argument is sensitive to how much emissions reductions affected the emergence of legal regulation. A state whose failure to reduce emissions significantly hindered the emergence of legal regulation would be more liable than a state whose equivalent failure to mitigate was politically inconsequential. Taking this more nuanced perspective adds some complexity but, importantly, does not fundamentally alter the way we typically think about liability for historical emissions. Bigger emitters would still be more liable than small emitters since, other things being equal, their emissions are more consequential for

the prospects for international cooperation. For example, it is more detrimental to international cooperation when China refuses to curb emissions than when Poland does so, simply because China's emissions play a larger role for climate change. Since the promotion effects will generally vary with the size of a state's emissions, we could probably approximate the promotion argument reasonably well simply by tracking contributions to cumulative emissions. It is worth recalling in this context, though, that the promotion argument only requires expending *reasonable* efforts. This means that the argument can explain the widespread intuition that states are not liable for so-called subsistence emissions (Shue 1993). States are not liable for subsistence emissions since refraining from such emissions would be unreasonably demanding. This reduces liability for emissions especially among poor states.

The promotion effects of emissions show why the familiar objection to promotion duties—that they are too vague to sustain ascriptions of liability—fails in at least this context. While there may be reasonable disagreement about exactly what different states were required to do fully to discharge their duty to promote just legal regulation, given the actual history of international climate politics, failing to significantly reduce non-subsistence emissions is enough to *violate* this duty. This is important because it shows that even if one were to doubt my idea that the extent of non-subsistence emissions yields a reasonably close approximation of states' failure to promote, emissions are not, as LLE/LLE* holds, free from liability just because they were not in breach of positive law. When a state's emissions amount to a violation of the duty to promote, the state is liable for those emissions even if it might well have been unclear whether the emissions were excessive in some more substantive sense.

An objection to the promotion argument is that the link between curbing emissions and promoting legal regulation is more tenuous than I have been letting on since a state often may make legal regulation *more* likely by emitting. This is the case, so this objection goes, whenever a state's emissions help create the need for legal regulation. But making legal regulation more likely is not sufficient for discharging one's duty to promote. After all, when a burglar 'incentivizes' homeowners to form a neighborhood watch, we do not think that he is thereby promoting greater home security. It is crucial to consider the intentions of the action, that is, whether the action is performed *in order to* discharge one's promotion duty. That is why a state like the United States should not be seen as having worked for an international climate treaty simply because its emissions helped create the need for it. Doing so would blur the distinction between the burglar and the institutions created to contain the burglar.

But *what* did states have a duty to work towards? The United States, to keep using that example, did not oppose international climate agreements across the board.

Instead, glossing over some detail, it argued that a fair agreement must also regulate developing-country emissions (Agrawal & Andresen 1999). While the country was guilty of undermining specifically the Kyoto Protocol, what is to say that they did not discharge their promotion duties in some other way? This objection necessitates adding some further nuance to what we have said so far. A first thing to note is that it is not enough merely to have promoted *some* legal regulation. As Rawls put it, promotion duties require working for '*just* arrangements not yet established' (Rawls 1999, 99. Italics added). But how should we understand 'just' regulation here? The appropriate response is not to insist on a detailed answer as to how, say, the global carbon budget is to be shared between states. That would fall into the trap of presupposing a nonexistent or epistemically inaccessible standard and thus be vulnerable to the critics discussed in this paper. Instead, we should index the promotion duties against a fairly minimal conception of legal regulation which allows for reasonable disagreement. A plausible proposal is that a relevantly just treaty, judged from a pre-legal perspective, is one that at least (i) curbs global emissions and (ii) allocates emissions more evenly than the status quo. If this is correct, any state that without unreasonable sacrifice could have made a treaty fulfilling (i) and (ii) more likely was under a duty to do so. It is plausible that many states failed to meet this condition, either because they did not work for such a treaty or because they did not make reasonable sacrifices in order to bring it about. Hence, the condition is sufficient to ground liability in many states.

Some may have lingering doubts about the promotion argument because we tend to be uneasy about retroactive legal liability. The promotion argument says that it may well be permissible for a climate treaty to hold states legally liable for emissions that, at the time of emitting, were in breach of no legally promulgated duty. To these readers who feel that this is a problem, I would like to offer three concluding thoughts. First, in holding states liable for pre-legal emissions, we are not holding them liable according to a standard that did not exist at the time of emitting. The promotion argument holds that states are liable for pre-legal emissions insofar as the emissions breached a duty that they *did* have—namely, the duty to promote. Thus, the argument does not rely on holding states under moral duties that emerged only later in time. Second, while the promotion argument maintains that breaching such duties is grounds for retroactive legal liability, it is important to stress that the issue at hand is not about punishment. In holding states liable for past emissions, we are not inflicting morally condemnatory harm as much as we are seeking to distribute the responsibility for dealing with a common problem. This is significant because while resistance to retroactivity is very strong in relation to criminal law—indeed, even a human right—we may find retroactivity less controversial in other

areas of law, including international treaty law.[23] It is not uncommon, for example, for governments to implement taxation law retroactively (Gribnau & Pauwels 2013).[24] Third, the main reason retroactivity in law is considered a problem is that it gives actors insufficient control over whether they are held liable: actors should receive a fair warning about which conduct is prohibited so that they have a chance to adjust their behavior accordingly (Hart 2008).[25] But actors can receive a 'fair warning' that some conduct is liable to future retroactive regulation even though positive law is yet to be enacted. In the climate case at least, states were surely put on notice that emissions were likely to be taken as grounds for liability at least since 1990. In terms of giving fair warnings, then, it is plausible that liability for emissions emerged well before the enactment of legally binding climate treaties.

6. Conclusion

Critics have argued that actors are not liable for greenhouse gas emissions prior to legitimate laws regulating emissions. I have challenged this argument by suggesting that emissions might run afoul of the natural duty to promote the *emergence* of such laws. To be sure, this does not show that the costs of climate policy ought to be allocated simply as a function of past emissions, as straightforward versions of the Polluter Pays Principle suggest. The promotion argument attenuates the relationship between emissions and liability and necessitates paying attention to precisely how emissions got (or get) in the way of effective political solutions to climate change. Yet the argument shows that even if one accepts premises like those endorsed by Pavel and Bou-Habib, there is nothing unfair about being held liable for pre-legal emissions when the emissions served to prevent just legal regulation of climate change, the polluter was aware of this, and the emissions were reasonably avoidable.

---

[23] Article 11 of the Universal Declaration of Human Rights holds that, 'No one shall be held guilty of any penal offence on account of any act or omission which did not constitute a penal offence, under national or international law, at the time when it was committed.' This corresponds to the legal maxim *nulla poene sine lege*, which traditionally includes three adjacent principles: (1) no punishment unless for a crime; (2) no action is a crime unless it violates the law; (3) no retroactive application of criminal law. For a seminal treatment of *nulla*, see Hall (1937); see also Popple (1989).

[24] My hunch is that this difference is explained by criminal law's condemnatory function since, in terms of bare setbacks to interests, the consequences of changing a tax code can be greater than a criminal fine. However, Robert Goodin has in conversation suggested that the key difference is that people can *comply with* retroactive tax laws—they can simply pay the new tax—whereas this is not the case for crimes already committed. In the United States, there is a specific debate about whether the constitution prohibits retroactive laws generally or only retroactive criminal law (Zoldan 2015).

[25] Another common complaint about retroactivity in law is that it allows the legislator to pick (with the benefit of hindsight) 'winners' and 'losers', thus enabling favoritism and corruption (Bell 1999).

References

Agrawala, Shardul, and Steinar Andresen. 1999. "Indispensability and Indefensibility? The United States in the Climate Treaty Negotiations." *Global Governance* 5 (4): 457–82.

Ashford, Elizabeth. 2003. "The Demandingness of Scanlon's Contractualism." *Ethics* 113 (2): 273–302.

Barrett, Scott. 2007. *Why Cooperate? The Incentive to Supply Global Public Goods*. Oxford: Oxford University Press.

Bell, Bernard. 1999. "In Defense of Retroactive Laws." *Texas Law Review* 78(1): 235–268.

Bell, Derek. 2011a. "Global Climate Justice, Historic Emissions, and Excusable Ignorance." *The Monist* 94 (3): 391–411.

Bell, Derek. 2011b. "Does Anthropogenic Climate Change Violate Human Rights?" *Critical Review of International Social and Political Philosophy* 14(2): 99–124.

Blomfield, Megan. 2016. "Historical Use of the Climate Sink." *Res Publica* 22 (1): 67–81.

Boden, T., R. Andres, and G. Marland. 2017. "Global, Regional, and National Fossil-Fuel CO2 Emissions (1751 - 2014) (V. 2017)." Environmental System Science Data Infrastructure for a Virtual Ecosystem; Carbon Dioxide Information Analysis Center (CDIAC), Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States).

Bou-Habib, Paul. 2019. "Climate Justice and Historical Responsibility." *The Journal of Politics*. Online first: 1–35.

Bowman, Paul. 2019. "On the Alleged Insufficiency of the Polluter Pays Principle." In *Studies on Climate Ethics and Future Generations. Vol I*, edited by Bowman and Berndt-Rasmussen. Institute for Futures Studies Working Papers. Stockholm: Institute for Futures Studies.

Buchanan, Allen, and Robert O. Keohane. 2006. "The Legitimacy of Global Governance Institutions." *Ethics & International Affairs* 20 (4): 405–37.

Caney, Simon. 2005. "Cosmopolitan Justice, Responsibility, and Global Climate Change." *Leiden Journal of International Law* 18 (4): 747–75.

———. 2006. "Environmental Degradation, Reparations, and the Moral Significance of History." *Journal of Social Philosophy* 37 (3): 464–82.

———. 2010. "Climate Change and the Duties of the Advantaged." *Critical Review of International Social and Political Philosophy* 13 (1): 203–28.

Cripps, Elizabeth. 2013. *Climate Change and the Moral Agent: Individual Duties in an Interdependent World*. Oxford: Oxford University Press.

Daniel Farber. 2017. "How Legal Systems Deal with Issues of Responsibility for Past Harmful Behavior." In *Climate Justice and Historical Emissions*, edited by Lukas H. Meyer and Pranay Sanklecha, 80–106. Cambridge: Cambridge University Press.

Duus-Otterström, Göran. 2014. "The Problem of Past Emissions and Intergenerational Debts." *Critical Review of International Social and Political Philosophy* 17 (4): 448–69.

Feinberg, Joel. 1974. "Noncomparative Justice." *The Philosophical Review* 83 (3): 297–338.

Gardiner, Stephen M. 2011. *A Perfect Moral Storm: The Ethical Tragedy of Climate Change*. Oxford: Oxford University Press.

Gartzke, Erik, Shannon Carcelli, Andres Gannon & Jack Zhang. 2017. "Signalling in Foreign Policy". *Oxford Research Encyclopedia of Politics.*

Goodin, Robert E. 1994. "Selling Environmental Indulgences." *Kyklos* 47 (4): 573–96.

———. 2013. "Disgorging the Fruits of Historical Wrongdoing." *American Political Science Review* 107 (3): 478–91.

Gribnau, Hans, & Pauwels, Melwin (Eds.) (2013). *Retroactivity of Tax Legislation*. Amsterdam: European Association of Tax Law Professors.

Hall, Jerome. 1937. "Nulla Poena Sine Lege." *The Yale Law Journal* 47 (2): 165–93.

Hart, H. L. A. 2008. *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford: Oxford University Press.

Keohane, Robert and David Victor. 2011. "The Regime Complex for Climate Change." *Perspectives on Politics* 9 (1): 7–23.

Klosko, George. 2004. *The Principle of Fairness and Political Obligation*. Rowman & Littlefield Publishers.

Marmor, Andrei (2011). *Philosophy of Law.* Princeton: Princeton University Press.

Meyer, Lukas H., and Dominic Roser. 2010. "Climate Justice and Historical Emissions." *Critical Review of International Social and Political Philosophy* 13 (1): 229–53.

Miller, David. 2007. *National Responsibility and Global Justice*. Oxford: Oxford University Press.

Miller, David. 2009. "Global Justice and Climate Change: How Should Responsibilities Be Distributed?" *Tanner Lectures on Human Values* 28: 119–56.

Page, Edward A. 2011. "Climatic Justice and the Fair Distribution of Atmospheric Burdens: A Conjunctive Account." *The Monist* 94 (3): 412–32.

———. 2012. "Give It up for Climate Change: A Defence of the Beneficiary Pays Principle." *International Theory* 4 (02): 300–330.

Parker, Charles & Karlsson, Christer. 2010. "Climate Change and the European Union's Leadership Moment: an Inconvenient Truth?" *Journal of Common Market Studies* 48 (4): 923–943.

Pavel, Carmen. 2016. "A Legal Conventionalist Approach to Pollution." *Law and Philosophy* 35 (4): 337–63.

Popple, James. 1989. "The Right to Protection from Retroactive Criminal Law." *Criminal Law Journal* 13 (4): 251–262.

Rawls, John. 1999. *A Theory of Justice*. Revised. Oxford: Oxford University Press.

———. 2005. *Political Liberalism*. Ney York: Columbia University Press.

Sadeleer, Nicolas de. 2002. *Environmental Principles: From Political Slogans to Legal Rules*. Oxford: Oxford University Press.

Shue, Henry. 1993. "Subsistence Emissions and Luxury Emissions Symposium: Above the Boundaries: Ozone Depletion, Equity, and Climate Change." *Law & Policy* 15: 39–60.

———. 1999. "Global Environment and International Inequality." *International Affairs* 75 (3): 531–45.

———. 2011. "Face Reality? After You!—A Call for Leadership on Climate Change." *Ethics & International Affairs* 25 (1): 17–26.

Sunstein, Cass R. 2007. "Of Montreal and Kyoto: A Tale of Two Protocols." *Harvard Environmental Law Review* 31: 1–66.

Valentini, Laura. 2017. "The natural duty of justice in non-ideal circumstances: on the moral demands of institution-building and reform." *European Journal of Political Theory*. Online first.

Vanderheiden, Steve. 2008. *Atmospheric Justice: A Political Theory of Climate Change*. Oxford: Oxford University Press.

Victor, David G. 2011. *Global Warming Gridlock: Creating More Effective Strategies for Protecting the Planet*. Cambridge: Cambridge University Press.

Walsh, James. 2007. "Do States Play Signaling Games?" *Cooperation and Conflict* 42 (4): 441–459.

Zoldan, Evan. 2015. "The Civil Ex Post Facto Clause" *Wisconsin Law Review* 4: 727–784.

Paul Bowman[1]

Duties of Corrective Justice and Historical Emissions[2]

This paper addresses the question of whether agents have incurred duties of corrective justice to bear the costs of climate change in virtue of having produced historical emissions, or emissions produced when it was still reasonable to be ignorant of the causes and harmful consequences of climate change. It argues that it is likely that agents have incurred duties of corrective justice in virtue of having produced some of their historical emissions, given that it is likely that they would have produced these emissions had they known, when they produced them, that the emissions would contribute to harmful climate change.

[1] Institute for Futures Studies & Department of Philosophy, Stockholm University, paul.bowman@iffs.se.
[2] Financial support by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences) is gratefully acknowledged.

1. Introduction

It is highly plausible that at least some agents—individual persons and perhaps entities like states and corporations (if they are indeed independent agents)—have moral duties to bear some of the costs of addressing climate change.[3] A number of philosophers have argued that these duties are justified, at least in part, by the principle of corrective justice—roughly, the principle that agents who wrongfully cause or contribute to a harm or a threat of future harm incur duties to bear the costs of rectifying the harm or threat.[4] According to these philosophers, because climate change has likely already resulted in harm and has the potential to result in future harm, and because many agents have wrongfully contributed to climate change by emitting excessive quantities of greenhouse gases, these agents have incurred duties of corrective justice to address climate change by bearing at least some of the costs of doing so.[5]

Despite the apparent simplicity of the above argument, the application of the principle of corrective justice to the circumstances of climate change faces several difficult questions. One question that has generated considerable debate among theorists concerns whether agents have incurred duties of corrective justice in virtue of having produced so-called "historical emissions," or emissions produced when it was still reasonable for agents to be ignorant of the causes and harmful consequences of climate change. This is the question that I will address in this paper.

There is some uncertainty about when it was still reasonable for emitters to be ignorant of the causes and harmful consequences of climate change. Following several others, I will assume that historical emissions are emissions produced prior to 1990, the year the Intergovernmental Panel on Climate Change (IPCC) published its First Assessment Report.[6] I will likewise refer to emissions produced after 1990 as "non-historical emissions."

---

[3] According to the standard categorization, these costs include mitigation costs (the costs of reducing greenhouse gas emissions), adaptation costs (the costs of protecting people from the adverse impacts of climate change), and compensation costs (the costs of compensating people for the adverse impacts of climate change). See, e.g., Caney, 2012.

[4] See, e.g., Caney, 2005; Vanderheiden, 2008; Bell, 2010; and Cripps, 2013. Defenders of the corrective justice approach tend to defend different versions of the principle. I avoid taking a stand on which version of the principle is correct, apart from my central concern, which is the relevance of reasonable ignorance to duties of corrective justice. Note also that I am conceiving of corrective justice somewhat narrowly. Some theorists take a wider view of corrective justice and hold that principles of corrective justice include principles that assign duties to bear costs in virtue of having benefitted from wrongful behavior. I take no stance on whether these duties are appropriately called "duties of corrective justice" or whether there are such duties. I am only interested in principles that assign duties to those who have themselves engaged in wrongful behavior.

[5] The principle of corrective justice, as interpreted and applied to the circumstances of climate change, is typically referred to as "the polluter pays principle." See, e.g., Caney, 2005.

[6] Among those who place the date at or around 1990 include: Singer, 2002; Caney, 2005; and Vanderheiden, 2011.

Several philosophers have argued that agents have not incurred duties of corrective justice in virtue of having produced historical emissions.[7] According to these philosophers, corrective justice does not require that agents bear the costs of rectifying the harmful consequences of their actions when the agents were reasonably (i.e., non-culpably) ignorant that their actions would have these consequences.

In this paper, I argue that these philosophers are wrong. I argue that it is likely that many agents have incurred duties of corrective justice to bear some of the costs of addressing climate change in virtue of having produced at least some of their historical emissions.[8] The basic argument I advance is relatively simple, though, as we will see, there are complications to work through. Roughly, I argue that: (1) in general, agents who were reasonably ignorant that their actions would have harmful effects do not avoid incurring duties of corrective justice if they would have performed the same (or a relevantly similar) action had they known about these effects, and (2) it is likely that many agents who produced excessive historical emissions would have produced these emissions even had they known, when they produced them, that the emissions would contribute to harmful climate change, as evidenced by their failure to reduce their emissions when they actually became aware of these effects. Although this basic line of argument has been previously suggested in the literature discussing historical emissions, it has not, to my knowledge, been developed in any detail.[9] In what follows, I develop and defend the argument.

It is important to note that even if I am wrong, and agents have not incurred duties of corrective justice in virtue of having produced historical emissions, this does not mean that these agents lack duties of corrective justice altogether. This is because most agents who produced excessive historical emissions also produced excessive non-historical emissions. If the corrective justice approach to climate justice is generally sound, then these agents have incurred duties of corrective justice in virtue of having produced excessive non-historical emissions.

---

[7] Caney, 2005, pp. 761-2; Vanderheiden, 2011, Ch. 6; Risse, 2008; Wündisch, 2017; Schüssler, 2011. Note that in a later article, Caney argues that emitters are liable for their historical emissions, but only to the extent that they have benefitted from these emissions. Caney, 2010, 203-228. For a similar view, see Bell, 2011, 391–411.

[8] Other philosophers, including Henry Shue (1999), Stephen Gardiner (2011), and Alexa Zelletin (2015) have also argued that agents have incurred duties of corrective justice to bear costs in virtue of having produced historical emissions. My argument is significantly different from each of theirs. Whereas Shue and Gardiner argue for a strict liability (or "no fault") approach to assigning duties of corrective justice, Zelletin argues that states are liable for their historical emissions in virtue of failing to investigate the effects of their emissions.

[9] Simon Caney (2010) mentions the argument but quickly rejects it. In a footnote to his discussion, Caney thanks Andrew Williams for pressing this line of argument during Caney's presentation of the paper (Caney, 2010, p. 209, fn. 14). Jonathan Pickering and Christian Barry (2012, p. 674) also very briefly suggest a similar line of argument.

This is not to say that the debate about historical emissions is inconsequential. Whether agents have incurred duties of corrective justice in virtue of having produced historical emissions is highly relevant to the overall *extent* of an agent's duty of corrective justice, or the overall cost of addressing climate change that the agent is required to bear. It is a complex question precisely how liability for historical emissions would affect the extent of different agents' duties, but it is a safe bet that agents who produced large quantities of historical emissions would have considerably more extensive duties of corrective justice than they otherwise would.[10] This is because it is highly plausible that, other things being equal, the extent of an agent's duty of corrective justice to rectify a harm or threat is proportional to the extent of the agent's wrongful causal contributions to the harm or threat.[11]

Notice that the question of whether agents have incurred duties of corrective justice in virtue of having produced historical emissions is much more consequential if entities like states and corporations can incur duties of corrective justice to bear costs associated with climate change. This is because the total amount of historical emissions attributable to the activities of these entities far exceeds the total amount of historical emissions attributable to the actions of living individual persons, simply due to the former's lack of a natural lifespan. However, the question of whether entities like states and corporations are agents that can incur duties, and especially those that arise from activities that were performed in the distant past, is controversial.[12] Nevertheless, given the focus on the obligations of states in international climate policy, I will assume that states are among the appropriate bearers of duties of corrective justice. Moreover, I will assume that the emissions produced within a state's borders are causally attributable to the state's policy choices.[13] Finally, because most historical emissions were produced in rich, developed states (like the United States), my discussion will focus on the policies and duties of these states.

One final note. I am only concerned with whether agents *have incurred* (or are likely to have incurred) duties of corrective justice in virtue of having produced historical emissions. It is a separate question, I take it, whether agents should be *held* liable, or asked (or forced) to fulfill the duties they have incurred. For example, there may be procedural or pragmatic reasons that strongly count against holding agents liable for their historical emissions. I will ignore these reasons and focus on

[10] Virtually all theorists of climate justice appear to accept this view. See, e.g.: Caney, 2005.
[11] Yet for a view skeptical of this claim, see: Tadros, 2018.
[12] For a discussion of some of the relevant issues in this debate, see: Stilz, 2011, and Boonin, 2011, Chapters 2 and 3.
[13] Note that this does not rule out that other agents (individuals, corporations, and even other states) can also be causally and morally responsible (and hence liable) for these emissions as well.

the prior question of whether agents have incurred duties of corrective justice in virtue of having produced historical emissions.

The basic plan for the remainder of the paper is this. In Section II, I will discuss the principle that, from the perspective of corrective justice, reasonable ignorance is fully exculpatory: an agent does not incur a duty of corrective justice to bear the costs of rectifying the harmful consequences of one's action when one was reasonably ignorant that one's action would have those consequences. In Section III, I will present my argument against the principle that reasonable ignorance is fully exculpatory. I will argue that reasonable ignorance is not exculpatory when the agent would have performed the same (or a relevantly similar) action under sufficiently better epistemic circumstances. In Section IV, I argue that, prior to 1990, it is likely that rich, developed states would not have substantially reduced their emissions had they been aware that their emissions would contribute to harmful climate change. The main evidence for this claim is the fact that these states failed to reduce their emissions when they actually became aware that their emissions would have this effect. In Sections V and VI, I will present and respond to two important challenges to the inference from what states have done to what they would have done at an earlier time. I briefly conclude in Section VII.

2. The Putative Basis of the Reasonable Ignorance Excuse

My aim in what follows is to investigate whether agents have incurred duties of corrective justice in virtue of having produced historical emissions. Notice, however, that even if reasonable ignorance is not exculpatory, there may be other reasons for why agents have not incurred duties of corrective justice in virtue of having produced at least some of their historical emissions (e.g., other justifications or excuses that agents may have). Therefore, I will only consider whether agents have incurred duties of corrective justice in virtue of having produced historical emissions in circumstance that are relevantly like the circumstances in which agents have incurred duties of corrective justice in virtue of having produced non-historical emissions.

I will therefore also assume that many agents have incurred duties of corrective justice to bear some of the costs of rectifying harmful climate change in virtue of having produced some non-historical emissions. Specifically, I will assume that an agent has incurred a duty of corrective justice in virtue of having produced emissions when:

(i) the agent's production of the emissions was all-things-considered, fact-relative morally wrong,[14]

(ii) at the time the agent produced the emissions, the agent believed, or ought to have believed, given the evidence available to the agent, that these emissions were contributing to harmful climate change (i.e., they were produced after 1990), and

(iii) the agent produced the emissions in the absence of any other conditions that might defeat or diminish to a significant degree the agent's responsible agency.

To simplify my discussion, "non-historical emissions" will henceforth refer to emissions in which an agent's production of those emissions satisfies conditions (i), (ii), and (iii). Likewise, "historical emissions" will henceforth refer to emissions in which an agent's production of those emissions satisfies conditions (i) and (iii), but in which, at the time the agent produced the emissions, the agent was reasonably ignorant that the emissions would contribute to harmful climate change (i.e., they were produced prior to 1990).

Given my assumption that agents have incurred duties of corrective justice by producing non-historical emissions, I will assume, more generally, that an agent, A, incurs a duty of corrective justice to bear some of the costs of rectifying a harm or threat of harm, h, in virtue of performing an action, ϕ, when:

(a) A's ϕ-ing causes or contributes to h,

(b) it is all-things-considered, fact-relative morally wrong for A to cause or contribute to h by ϕ-ing,

(c) at the time of A's ϕ-ing, A believes, or ought to believe given the evidence available to A, that ϕ-ing would cause or contribute to h, and

(d) A ϕs in the absence of any other conditions that might defeat or significantly diminish A's responsible agency.

[14] The notion of fact-relative wrongness comes from Derek Parfit (2011). According to Parfit, an action is fact-relative wrong if it would be wrong for a person who knows all the facts to perform the action (2011, pp. 150-1).

When an agent performs an action that satisfies conditions (a), (b), (c), and (d), I will call the agent a 'culpable causer.'

To investigate whether agents have incurred duties of corrective justice by producing historical emissions, I will therefore consider the question of whether an agent incurs a duty of corrective justice to rectify a harm or threat of harm by performing an action that satisfies (a), (b), and (d) but in which, at the time that the agent performs the action, the agent is reasonably ignorant that the action would cause or contribute to the harm or threat of harm. I will call such an agent an 'ignorant causer.'

I will assume, then, that the following principle underlies the claim that agents have not incurred duties of corrective justice in virtue of having produced historical emissions:

Reasonable Ignorance is Exculpatory (RIE):

If:

(a) an agent A's ϕ-ing causes or contributes to a harm or threat of harm, h,

(b) it is all-things-considered, fact-relative morally wrong for A to cause or contribute to h by ϕ-ing,

(c) at the time of A's ϕ-ing, A is reasonably ignorant that A's ϕ-ing would cause or contribute to h, and

(d) A ϕs in the absence of any other conditions that might defeat or significantly diminish A's responsible agency,

then:

(e) A does not incur a duty of corrective justice to bear any of the costs of rectifying h (i.e., preventing or compensating for h) in virtue of ϕ-ing.

In short, (RIE) states that an ignorant causer does not incur a duty of corrective justice to bear any of the costs of rectifying the harm or threat of harm she causes or contributes to causing.

In what follows, I aim to demonstrate not only that (RIE) is false, but that (RIE)

is false in a way that implies that agents have incurred duties of corrective justice in virtue of having produced historical emissions. To give an example of an ignorant causer, and to illustrate (RIE), consider the following case, which is closely adapted from a case created by Thomson (who uses the case for a somewhat different purpose):

> *Light Switch*: Bronn always comes home at 9:00PM, and the first thing he does is to flip the light switch in the hallway. He does so this evening. Bronn's flipping the switch causes a circuit to close. By virtue of an extraordinary series of coincidences, unpredictable in advance by anybody, the circuit's being closed causes a release of electricity (a small lightning flash) in V's house next door. Unluckily, V is in its path and is therefore badly burned.[15]

Bronn is an ignorant causer. Bronn's flipping the switch both causes a serious harm to V and is all-things-considered, fact-relative morally wrong. Additionally, at the time Bronn flips the switch, Bronn is reasonably ignorant that his flipping the switch would cause the harm to V. Therefore, (RIE) implies that Bronn does not incur a duty of corrective justice to compensate V. (Note also that I am assuming that had Bronn believed that his flipping the switch would cause the harm to V, then B would have incurred a duty of corrective justice to compensate V.)

What evidence is there for (RIE)? Philosophers who have argued that agents have not incurred duties of corrective justice in virtue of having produced historical emissions have typically argued for a principle like (RIE) by claiming that it would be *unfair* for an agent to bear the costs of rectifying a harm or threat of harm that the agent did not know, and could not have reasonably known, that she would cause or contribute to causing.[16] However, these philosophers have not fully spelled out the argument for precisely why it would be unfair for an ignorant causer to bear these costs.[17] It is important to get clear on this argument, especially because some philosophers who have argued that agents *have* incurred duties of corrective justice in virtue of having produced historical emissions have rejected a principle like (RIE) by claiming that it would be unfair for an ignorant causer *not* to bear the costs of rectifying the harm or threat of harm that she causes or contributes to causing

[15] Thomson, 1990, p. 229.
[16] See, e.g.: Caney, 2005, p. 762; Vanderheiden, 2008, Ch. 6; Wündisch, 2017.
[17] Caney (2005, p.762), for instance, appeals to a distinction between what he calls "the perspective of the duty-bearers" versus "the perspective of the rights-holders," and claims that holding agents liable in virtue of their having produced historical emissions would be to unfairly prioritize the perspective of the rights-holders over the perspective of the duty-bearers. However, Caney does not explain *why* doing so would unfairly prioritize the perspective of the rights-holders over that of the duty-bearers.

(again, without very much argument).[18] Therefore, I will attempt to spell out a case for (RIE) that is based on the purported unfairness of requiring that an ignorant causer bear the costs of rectifying the harm or threat of harm that she causes or contributes to causing. For simplicity, I will focus on cases of a single ignorant causer, i.e., where no other agent is causally responsible for the harm or threat. Additionally, I will focus on cases in which the harm has already occurred, as opposed to one in which the harm is threatened. Nothing that might undermine my argument rests on these choices.

In general, it is common for philosophers to hold that considerations of fairness (or justice) determine whether and to what extent agents incur duties of corrective justice.[19] The basic idea is that corrective justice is ultimately concerned with fairness in the distribution of the costs of wrongful behavior. With respect to wrongful behavior that causes harm to others, corrective justice considers whether and to what extent it would be fair for the agent who wrongfully causes the harm to bear the cost of the harm (by rectifying it), rather than for those who did not cause the harm to bear the cost of the harm. If it would be fair for the agent to bear at least a portion of the cost of the harm she causes, then the agent incurs a duty of corrective justice to bear that cost in service of rectifying the harm.[20] If it would not be fair for the agent to bear even a portion of the cost of the harm she causes, then the agent does not incur a duty of corrective justice to bear any of the cost of the harm. Note, finally, that if the cost that an agent incurs a duty of corrective justice to bear is less than the total cost of the harm that the agent causes, then whether it is fair for the *victim* of the harm to bear the remaining cost herself (by, e.g., suffering the uncompensated harm) is determined by considerations other than those of corrective justice (like considerations of, say, distributive justice or beneficence).

Therefore, if it would be fair for an ignorant causer to bear some portion of the cost of the harm she causes, rather than for others to bear the entire cost, then the ignorant causer incurs a duty of corrective justice to bear that portion of the cost, and (RIE) is false. If, on the other hand, it would not be fair for the ignorant causer to bear even a portion of the cost of the harm, then the ignorant causer does not incur a duty of corrective justice to bear any of the cost of rectifying the harm, and (RIE) is true.

Why, then, might it not be fair for the ignorant causer to bear even a portion of the cost of the harm she causes? To answer this question, it is important first to note,

---

[18] See Shue, 1999, p. 535-6 and Gardiner, 2011, Ch. 11.

[19] See, e.g., Coleman, 1995, and Tadros, 2011.

[20] Note that the question that is relevant to corrective justice is whether and to what extent it is fair for the causer of the harm *qua causer of the harm* to bear the cost of the harm. It may be fair for the causer of the harm to bear a portion of the cost of the harm on the basis of, say, her ability to bear it. But if the causer of the harm incurs a duty to bear the cost for this reason, the duty is not one of corrective justice.

as McMahan and others have pointed out, that there is a general moral presumption against shifting harm, including the cost of a harm.[21] That is, morality does not require, and often prohibits, that individuals suffer the harms, or bear the cost of the harms, that befall others unless there is a morally significant reason for them to do so. For example, if a boulder naturally dislodges from a cliff and lands on my car, then under typical circumstances (e.g., in the absence of a preexisting liability agreement), you do not have a duty to compensate me for the damage the boulder causes, and I am not permitted to force you to compensate me. The bad luck is mine, and other things being equal, it would not be fair for you to bear the cost of my bad luck.

That there is a presumption against shifting harm does not mean, of course, that the presumption cannot be overcome. Perhaps, for example, considerations of beneficence can justify shifting some of the costs of particularly serious harms to those who can bear the costs without significant reductions to their well-being.[22] Moreover, with respect to corrective justice, nearly everyone believes that the presumption against shifting harm is overcome in cases in which the harm is caused by a *culpable* causer. Suppose, for example, that for no good reason you intentionally dislodge a boulder from a cliff with the goal of damaging my car, and you succeed in doing so. Your culpability for causing the harm is sufficient to justify shifting the entire cost of the harm to you. It is fair for you, rather than me (or any other non-responsible party), to bear the entire cost of the harm, and so you incur a duty of corrective justice to compensate me fully for the harm you wrongfully caused.

Given this presumption against shifting harm, the question, then, is whether the presumption is overcome in cases involving an ignorant causer. In other words, is there sufficient justification to shift even a portion of the cost of the harm caused by an ignorant causer from the victim of the harm *to* the ignorant causer—namely, a reason that makes it fair for the ignorant causer to bear at least a portion of the cost of the harm? According to defenders of (RIE), there is not. Therefore, the cost of the harm ultimately must be borne either by the victim or by third parties (perhaps according to some other distributive principle).

How might defenders of (RIE) attempt to justify the claim that there is not sufficient justification for shifting a portion of the cost of the harm to the ignorant causer? First, consider why, in the boulder-pushing case, it is fair for you (the culpable causer), rather than for me (the victim) or third parties, to bear the costs of rectifying the harm you caused. That is, why does a person's culpability for causing a harm make it fair for the culpable causer to bear the cost of the harm?

A plausible explanation appears to be lie, in large part at least, in the difference

[21] McMahan, 1994.
[22] See, e.g., Singer, 1972.

in the quality of the *opportunity* that each party had to avoid bearing the cost of the harm.[23] You (the culpable causer) had an excellent opportunity to avoid bearing the cost of the harm—that is, to avoid bearing the cost of compensating me. You could have avoided bearing this cost simply by choosing to refrain from causing the harm, which given your culpability, you were morally obligated to do and were fully capable of doing. Thus, you have, at best, only a very weak objection to bearing the full cost of the harm. On the other hand, through no fault of my own, my property is destroyed, and I had no reasonable way to avoid bearing this cost. Similarly, third parties had no reasonable way to avoid bearing the cost of the harm (if the cost were ultimately allocated to them). Hence I, as well as third parties, have a very powerful objection to bearing the cost of the harm. It would therefore be fair for you, rather than for me or any third parties, to bear any of the cost of the harm, and so you incur a duty of corrective justice to compensate me.

Notice, however, that a similar explanation cannot be given for why it would be fair for the ignorant causer, rather than for the victim or third parties, to bear the cost of a harm caused by the ignorant causer. The ignorant causer, like the victim and third parties, has a strong objection to bearing the cost of the harm in virtue of lacking a reasonable opportunity to avoid bearing it. For instance, in *Light Switch*, although V is in no way liable to bear the cost of the harm and lacked a reasonable opportunity to avoid suffering the harm, Bronn also lacked a reasonable opportunity to avoid bearing the cost of the harm. Because Bronn did not know that his flipping the switch would cause the harm, Bronn did not choose to cause the harm, either intentionally or as a side-effect of an intended outcome. So unlike the culpable causer, Bronn did not have a reasonable opportunity to avoid causing the harm. Therefore, unlike a culpable causer, Bronn has a powerful objection to bearing the cost of the harm.

The defender of (RIE) can claim that because the victim of the harm and the ignorant causer each has a strong objection to bearing the cost of the harm (in virtue of lacking a reasonable opportunity to avoid bearing this cost), there does not appear to be any moral basis on which to justify shifting the cost of the harm from the victim of the harm to the ignorant causer. Therefore, the presumption against shifting the cost of the harm to the ignorant causer is not overcome. According to the defender of (RIE), it would be fair for either the victim of the harm or third parties (on the basis of some other principle), rather than the ignorant causer, to bear the cost of the harm. Therefore, the ignorant causer does not incur a duty of corrective justice to bear any of the cost of the harm.

This argument for (RIE) is plausible. Nevertheless, in the next section, I will

[23] See, e.g., Scanlon, 1998, Ch. 6; Tadros, 2011, McMahan, 2002, p. 401.

argue that it fails to establish that it is *never* fair for an ignorant causer to bear the cost of the harm he causes. I will argue that there are circumstances in which it would be fair for an ignorant causer to bear the cost of the harm she causes, even though the ignorant causer lacked a reasonable opportunity to avoid bearing the cost of the harm (and despite the presumption against shifting the cost of the harm from the victim to the ignorant causer). Hence, these are circumstances in which an ignorant causer incurs a duty of corrective justice to bear the cost of the harm she causes. These circumstances, therefore, provide a counterexample to (RIE). Later, I will argue that these circumstances are also those in which many agents produced historical emissions.

3. An Exception to the Reasonable Ignorance Defense

In this section, I will argue that (RIE) is false. The claim that I will argue for, and that implies that (RIE) is false, is that an ignorant causer incurs a duty of corrective justice to bear the costs of rectifying a harm she causes if she *would have* performed the same action had she known, at the time she performed the action, that her action would cause the harm. More formally, I will argue for the following principle:

> *Exception to Reasonable Ignorance is Exculpatory* (ERIE): An agent, A, incurs a duty of corrective justice to bear some of the costs of rectifying a harm or threat of harm, h, in virtue of A's performing an action, ϕ, when:
>
> (a) A's ϕ-ing causes or contributes to h,
>
> (b) it is all-things-considered, fact-relative morally wrong for A to cause or contribute to h by ϕ-ing,
>
> (c) at the time of A's ϕ-ing, A is reasonably ignorant that A's ϕ-ing would cause or contribute to h,
>
> (d) A ϕs in the absence of any other conditions that might defeat or significantly diminish A's responsible agency, and

(e) had A known, at the time of A's ϕ-ing, that A's ϕ-ing would cause or contribute to h, A would have ϕ-ed.[24]

Recall that (RIE) holds that A does *not* incur a duty of corrective justice to bear any of the costs of rectifying a harm or threat of harm when conditions (a), (b), (c), and (d) are satisfied. So if (ERIE) is true, then (RIE) is false.

To illustrate (ERIE), consider again *Light Switch*. Recall that in *Light Switch*, (RIE) implies that, as an ignorant causer, Bronn does not incur a duty of corrective justice to compensate V. However, suppose that the following counterfactual is true of Bronn: had Bronn known, at the time that he flips the switch, that flipping the switch would cause the injury to V, Bronn would have flipped the switch anyway. We can add a few details to the case to provide further context for the counterfactual. Suppose that Bronn was motivated to flip the switch to find his way to the computer room so that he could execute a small trade (which he knows would prevent a $50 stock loss). Suppose, moreover, that at the time that he flips the switch, Bronn is completely indifferent to V's welfare, such that Bronn would have chosen to avoid any small cost to himself rather than refraining from causing even a serious harm to V. Therefore, had Bronn known that flipping the switch would cause harm to V, Bronn would have flipped the switch to avoid the small cost to himself. Call this variation of the case *Indifferent Switch*.

Here is why (ERIE) is true. If an ignorant causer would have performed the harm-causing action even had she known that her action would result in harm, then the ignorant causer acts *while* ignorant but does not act *from* ignorance.[25] That is, although the ignorant causer is, at the time that she performs the action, ignorant that her action would cause the harm, the ignorance does not play a role in explaining why the agent performs the action that she does. For example, in *Indifferent Switch*, Bronn's ignorance does not explain why Bronn flips the switch, given that he would have, had he not been ignorant, performed the same action and for the same reasons that he did (namely, to avoid the $50 stock loss).

According to defenders of (RIE), an ignorant causer's ignorance is supposed to explain why she does not incur a duty of corrective justice even though she performs an all-things-considered, fact-relative wrongful action that harms someone. After

[24] I believe that there is a defensible version (ERIE) in which condition (e) instead states that: had A known *or ought to have known*, at the time of A's ϕ-ing, that A's ϕ-ing would cause or contribute to h, A would have ϕ-ed. Investigating this possibility, however, will take me too far afield, and is unnecessary for establishing my main thesis.

[25] This distinction dates back to Aristotle (1999, Bk. 3, Ch. 1, 5). For other (brief) discussions of this distinction, see, e.g., Zimmerman, 1997, p. 424, and Peels, 2014, p.479.

all, a *culpable* causer, who is exactly like an ignorant causer except for the latter's ignorance, *does* incur a duty of corrective justice. But if an ignorant causer's ignorance does not explain why she performs the harm-causing action, then it is difficult to see how her ignorance can explain why she should be *exempt* from bearing the cost of the harm that results from her action (a cost that, had she not been reasonably ignorant, she would have incurred a duty to bear).

To consider this last claim further, recall why the defender of (RIE) holds that an ignorant causer's ignorance explains why she does not incur a duty to rectify the harm. According to the defender of (RIE), the ignorant causer could object to bearing the cost of the harm she causes by claiming that she lacked a reasonable opportunity to avoid bearing this cost in virtue of the fact that she did not choose to cause the harm, either intentionally or as a foreseen side-effect of an intended outcome. And while it is true that the victim could similarly object that she lacked a reasonable opportunity to avoid bearing the cost of the harm, because of the presumption against shifting the cost of the harm to the ignorant causer, it would be unfair for the ignorant causer to bear the cost of the harm. Therefore, the ignorant causer does not incur a duty of corrective justice to bear the cost of the harm.

Yet consider the ignorant causer's objection that she lacked a reasonable opportunity to avoid bearing the cost of the harm. The basis of the objection seems to be that if she were required to bear the cost of the harm, she would be the victim of bad luck. Given that she did not choose to cause the harm, either intentionally or as a foreseen side-effect, it is essentially a matter of bad luck that there is an unavoidable cost that someone must bear—no different, perhaps, than a harm caused by purely natural factors. And although it is unfortunate that someone must bear the cost of this bad luck, the cost is ultimately not hers to bear, given the presumption against shifting harm.

However, if the ignorant causer would have performed the same action even had she known that her action would cause the harm, then it is false that it is simply a matter of bad luck that there is an unavoidable cost that someone must bear. Thus, it is false that if the ignorant causer were required to bear the cost of the harm, she would be a victim of bad luck. Rather, if she were *not* required to bear the cost of the harm, she would be the beneficiary of very good luck at the expense of the victim's bad luck. For if by chance the ignorant causer had learned that her action would cause harm, then she would have performed the same action and would have been a culpable causer. And as a culpable causer, she would have incurred a duty to bear the cost of the harm, rather than the victim bearing this cost himself. Therefore, if we assume that the fair distribution of the cost of a harm should be sensitive to considerations of luck, then it is fair for the ignorant causer, rather than the victim, to bear the cost of the harm.

Even if we accept my argument for (ERIE), notice that the principle is quite narrow—it states that an ignorant causer incurs a duty to bear the cost of the harm she causes if she would have performed *the same action* had she known that her action would cause the harm. Can we say more than this? Suppose that, had the ignorant causer known that her action would cause harm, she would have performed an action that was morally *worse* (e.g., one that caused more harm overall) than the all-things-considered, fact-relative morally wrongful action she actually performs. For example, suppose that in *Light Switch*, had Bronn known that his flipping the switch would cause the injury to V, Bronn would not have flipped it. However, suppose that the reason that Bronn would not have flipped it is that Bronn intended to inflict a much greater injury on V than the one he actually inflicts by flipping the switch, and Bronn would not have been satisfied by the injury to V produced by flipping the switch. So had Bronn known that flipping the switch would have caused the injury to V, Bronn would not have flipped it, but would have instead opted to inflict an even greater injury on V.

It seems that if Bronn has a duty to compensate V in *Indifferent Switch*, then Bronn has a duty to compensate V in this variation as well. Although Bronn's ignorance explains why he flips the switch, his ignorance explains why he did not perform a morally worse action (rather than a morally permissible action). So Bronn has only a weak objection to bearing the cost of the harm he actually causes, given that had he known that his action would cause harm, he would have incurred a duty to bear an even greater cost (that is, the cost of rectifying the harm resulting from the morally worse action).

Now suppose that had an ignorant causer known that her action would cause harm, she would have performed an action that was morally *better* than the all-things-considered, fact-relative wrongful action she actually performs, yet one that was nevertheless all-things-considered morally wrong (e.g., one that would have caused harm, but just *less* harm than the action she actually performs). So suppose that had Bronn known that flipping the light switch would cause the injury to V, Bronn would not have flipped it, but would have instead inflicted a smaller harm on V. Suppose, for example, that prior to flipping the switch, Bronn intended to dump large amounts of trash into V's yard, and had Bronn not (unwittingly) caused the injury to V by flipping the switch, Bronn would have carried through on his initial trash-dumping plan.

In this case, it seems fair for Bronn to bear the cost of the (actual) harm he causes. His objection to bearing the cost of the harm is still relatively weak, given that he would have performed an all-things-considered wrongful action had he known that his action would cause harm. It is plausible, however, that the maximum cost that Bronn would be required to bear to rectify the harm is lower than it is in those

variations in which he would have performed the same (or a morally worse) action had he known that his action would cause harm.

Given my discussion, we can modify (ERIE) as follows:

> *Exception to Reasonable Ignorance is Exculpatory** (ERIE*): An agent, A, incurs a duty of corrective justice to bear some of the costs of rectifying a harm or threat of harm, h, in virtue of A's performing an action, ϕ, when:
>
> (a) A's ϕ-ing causes or contributes to h,
>
> (b) it is all-things-considered, fact-relative morally wrong for A to cause or contribute to h by ϕ-ing,
>
> (c) at the time of A's ϕ-ing, A is reasonably ignorant that A's ϕ-ing would cause or contribute to h,
>
> (d) A ϕs in the absence of any other conditions that might defeat or significantly diminish A's responsible agency, and
>
> (e) had A known, at the time of A's ϕ-ing, that A's ϕ-ing would cause or contribute to h, A would have performed an action that was all-things-considered morally wrong.

4. Applying (ERIE*)

The next step is to apply (ERIE*) to agents' production of historical emissions. Recall that I am assuming that agents' production of historical emissions satisfies conditions (a)-(d) of (ERIE*). I am also assuming that historical emissions only include emissions produced prior to 1990. Finally, I am assuming that states are agents that can incur duties of corrective justice to bear costs associated with climate change. The relevant question, then, is what states—and particularly rich, developed states—would have done, prior to 1990, had they known that their emissions would contribute to harmful climate change.

One important piece of evidence for what these states would have done, prior to 1990, had they known that their emissions would contribute to harmful climate change, is what states *actually* did, after 1990, when they knew that their emissions would contribute to harmful climate change. I have been assuming that after 1990, many states acted all-things-considered morally impermissibly by producing

excessive emissions. That is, I have been assuming that once states (or their leaders) became aware that their emissions would contribute to harmful climate change, these states produced more emissions than they were morally permitted to produce.

I do not think that this assumption is controversial, particularly for rich, developed states, which could have reduced their emissions substantially without sacrificing their citizens' high living standards. Consider that after 1990, not only did most states fail to reduce their emissions substantially, most states *increased* their emissions. [26] Overall, between 1990 and 2012, global annual emissions increased by an average of about two percent each year.[27] In the United States, annual emissions increased by an average of a quarter of a percent each year during that period.[28] And although some European countries did decrease their annual emissions between 1990 and 2012, these reductions were relatively small and were attributable, at least in part, to the outsourcing of manufactured goods consumed in these countries to developing countries.[29]

The fact that after 1990, many states, and particularly rich, developed states, did not reduce their emissions to morally permissible levels (or much at all) is good evidence that *prior* to 1990, had these states known that their emissions would contribute to harmful climate change, they would not have kept their emissions within morally permissible levels. In the absence of countervailing evidence, we are, I believe, entitled to conclude that it is likely that that prior to 1990, had these states known that their emissions would contribute to harmful climate change, they would not have kept their emissions within morally permissible levels. Therefore, given (ERIE*), we are entitled to conclude that it is likely that many states have incurred duties of corrective justice in virtue of having produced historical emissions.

What countervailing evidence, if any, do we have? Over the next two sections, I will consider and respond to two important challenges to the claim that states' failures to reduce their emissions to morally permissible levels after 1990 means that it is likely that prior to 1990, states would not have kept their emissions within morally permissible levels, had they known that their emissions would contribute to harmful climate change.

5. Challenge I: Path-Dependence

The first challenge considers the effect of *path-dependence* on states' lack of post-1990 action to reduce their emissions to morally permissible levels. It is highly

[26] WRI, 2014.
[27] WRI, 2014.
[28] WRI, 2014.
[29] WRI, 2014.

plausible that part of the explanation for why states are now, and have been since 1990, resistant to enact aggressive policies to reduce their emissions to morally permissible levels is that their economies are highly dependent on fossil fuels, making the costs of switching to low carbon energy sources very expensive. Yet when states first began to industrialize, their economies were not yet significantly dependent on fossil fuels. Perhaps it would have been considerably less costly, at that time, for states to enact policies that would have enabled them to develop economically, over the long term, using less carbon-intensive sources of energy. Therefore, perhaps had states known about the climate-altering potential of unabated fossil fuel use, they would have enacted these policies, and thus would have kept their emissions within morally permissible levels.

Let's suppose that, due to the effect of path-dependence, it is true that had states been given very early warning of the climate-altering dangers of their emissions, they would have enacted policies that would have kept their emissions within morally permissible levels.[30] Does this mean, then, that states' production of historical emissions does not satisfy condition (e) of (ERIE*)?

Notice that even if it is true that had states received very early warning of the climate-altering dangers of unabated fossil-fuel use, they would have enacted policies that would have kept their emissions within morally permissible levels, it might also be true—and, in fact, it is plausible—that had states come to learn about the causes and consequences of climate change at some later date, they would *not* have enacted policies that would have kept their emissions within morally permissible levels. With respect to the path-dependence of fossil fuel use, perhaps as states' economies became more dependent on fossil fuels, it became less likely that states would have changed course had they learned that their emissions would contribute to harmful climate change. For instance, suppose that the United States had learned, during the 1950's, that its emissions would contribute to harmful climate change. It seems likely that by this time, the United States' economy was dependent on fossil fuels to such an extent that the costs of decarbonization were

[30] I actually think that this claim is implausible. As Gardiner (2011) and others have persuasively argued, the lack of meaningful action on climate change by states since 1990 is explained by more than just the high current short-term economic costs of decarbonization, though this is obviously an important factor. States' lack of meaningful action on climate change is also explained by *who* will be primarily harmed by climate change, namely the global poor and future generations. Since past generations of policy-makers were probably equally (or perhaps even less) likely to consider the interests of the global poor and future generations than contemporary policy-makers are and have been, early warning of the climate-altering potential of greenhouse gas emissions would probably not have resulted in policies that ensured that emissions remained within morally permissible levels, even if the costs of enacting such policies were comparatively lower than they are today. Nevertheless, the worry about path-dependence is worth taking seriously if only because it blocks any easy inference from what states have done since 1990 to what states would have done prior to 1990 had they known that they would contribute to harmful climate change.

very high. If so, it is probably true that, had the United States learned, in the 1950's, that its emissions would contribute to harmful climate change, the United States would *not* have kept its emissions within morally permissible levels.

The issue that we must address, I believe, concerns how we should interpret the counterfactual that prior to 1990, had states known that their emissions would contribute to harmful climate change, states would not have reduced their emissions to morally permissible levels. This is because it appears that what states would have done depends on *when* they would have learned about the dangers of climate change. For the purposes of applying (ERIE*), should we consider what states would have done had they had very early warning of the causes and risks of climate change (in which case, perhaps states would have kept their emissions within morally permissible levels) or should we also consider what states would have done had they learned, at some later time or times, that their emissions would contribute to harmful climate change (in which case, perhaps states would not have kept their emissions within morally permissible levels)?

Let's consider the question in a simplified case. Consider:

> *Soup:* Kay discovers, through trial and error, a recipe for a delicious soup. The recipe requires rare ingredients, each of which must be added to the pot in a specific order. Completely unbeknownst to Kay, the specific combination of the ingredients causes the fumes of the soup to become extremely toxic during the cooking process (adding the final ingredient always detoxifies the soup, making it safe to eat). Fortunately for Kay, when cooking the soup, she always has her window open, which allows the fumes from the soup to escape her apartment without ever affecting her. Unfortunately, however, these fumes are blown into her next-door neighbor V's apartment. Kay chooses to make this soup several times, first at t_1, and then at later times, t_2, t_3, etc. After making soup at t_9, the cumulative effect of the fumes causes V to become very ill, which costs thousands of dollars to treat. At some point—t_{10}—Kay discovers that she is making V very ill. However, Kay continues to make the soup after this discovery. Kay's subsequent soup-making causes her neighbor to become extremely and permanently ill. The cost of treating her neighbor's illness is several thousand dollars more than the cost had been at t_{10}.

I assume that after t_{10}, Kay is a culpable causer, and so Kay has a duty to bear costs in virtue of her post-t_{10} choices to make the soup.

Yet Kay is an ignorant causer with respect to each of her choices to make soup prior to t_{10}. Moreover, suppose that the following counterfactual (C1) is true:

> C1: had Kay learned, prior to her first choice to make soup (at t_1), that making the soup would cause harm or risk causing harm to her neighbor, she would not have made the soup at any subsequent time.

Suppose, for instance, that prior to tasting the soup, Kay did not know how delicious the soup was. Having not known how delicious the soup was, she would have decided to put the interests of her neighbor over her desire to make the soup, and so she would not have made it.

Now suppose that when Kay actually makes the soup for the first time (at t_1), Kay discovers that the soup is far more delicious than she could have ever imagined. Suppose, moreover, that the following counterfactual is true:

> C2: had Kay learned, after t_1 but prior to t_2, that making the soup would cause or risk causing harm to V, Kay would have made the soup at t_2 and at all subsequent times that she actually made it.

C1 and C2 are consistent. We can assume, moreover, that Kay does not incur a duty of corrective justice to bear costs in virtue of her choice at t_1. Yet what should we say about whether Kay incurs a duty of corrective justice to bear costs in virtue of her subsequent choices to make soup, like her choice at t_2? According to C1, if Kay had known, prior to t_1, that making soup would cause harm, she would not have made soup at t_2. However, according to C2, had Kay learned that making soup would cause harm just after t_1, she would have made soup at t_2. Given (ERIE*), should Kay incur a duty to bear costs in virtue of her choice at t_2?

Here is an argument for why Kay does not incur a duty of corrective justice in virtue of her choice at t_2: Kay only causes her neighbor harm *at all* because of her ignorance prior to when she first made soup. Therefore, her ignorance explains all her subsequent choices to contribute to harm, since Kay would not have made these choices had she not been ignorant prior to t_1. Since this ignorance was reasonable, Kay should not incur a duty of corrective justice to bear costs in virtue of her choice at t_2.

This argument, however, is unsuccessful. First, notice that according to C1, had Kay learned, prior to t_1, that her soup-making would cause harm, Kay would not have chosen to make soup even after t_{10}, the time at which she *actually* discovered that her soup-making is causing harm. But clearly Kay does incur a duty to bear costs in virtue of her post-t_{10} choices, even if C1 is true. Therefore, the fact that Kay would not have chosen to make soup at t_2 had she known, prior to t_1, that her soup-making would cause harm does not necessarily show that she does not incur a duty in virtue of her choice at t_2.

To investigate further, we can distinguish between two kinds of effect that an agent's ignorance may have on the agent's choice. First, an agent's ignorance may affect the agent's choice by affecting the costs that the agent must bear to satisfy her preferences. Call this an *indirect effect*. C1 describes the indirect effect that Kay's ignorance, prior to t_1, has on Kay's choice to make soup at t_2 (as well as on all her subsequent choices, including those after t_{10}). We can infer, from C1, that Kay has a preference, at all times, not to cause her neighbor harm. Moreover, Kay's ignorance of the harmful consequences of her soup-making, prior to t_1, has the effect of raising the costs to her of satisfying this preference at t_2 (and at every subsequent time) in virtue of the fact that she forms a strong preference to make the soup only because she was ignorant of the harmful consequences of her soup-making. Thus, at t_2 (and every subsequent time), she has a stronger preference to make the soup than she has a preference to avoid causing her neighbor harm. Had Kay *not* been ignorant, prior to t_1, that making the soup would cause harm to her neighbor, it would have been relatively costless for Kay to satisfy her preference to refrain from causing her neighbor harm at t_2 (or any subsequent time), which she would have chosen to do.

An agent's reasonable ignorance is not exculpatory in virtue of its having an indirect effect on the agent's choice to perform a harm-causing action. For example, Kay's ignorance, prior to t_1, has an indirect effect on her choices to make the soup after t_{10}. However, her ignorance, prior to t_1, obviously does not defeat her duty to bear costs in virtue of her post-t_{10} choices. So even though Kay's ignorance, prior to t_1, has an indirect effect on her choice at t_2, her ignorance, prior to t_1, does not defeat her duty to bear costs in virtue of her choice at t_2, either.

An agent's reasonable ignorance is exculpatory only when it has what we can call a *direct effect* on the agent's choice. An agent's ignorance has a direct effect on the agent's choice when the ignorance affects whether the agent chooses the outcome that reflects her preferences at that time. C2 describes the *lack* of a direct effect that Kay's ignorance, at t_2, has on Kay's choice at t_2. To see this, suppose that C2 is false. Suppose that had Kay learned, after t_1 but prior to t_2, that making the soup would cause harm to V, Kay would not have made the soup at t_2. We can infer that at t_2, Kay has a stronger preference not to cause harm to her neighbor than she has a preference to make the soup. But K's ignorance, at t_2, prevents Kay from choosing the option that reflects these preferences, which is why she makes the soup at t_2.

Now suppose that C2 is true. We can infer that, at t_2, Kay has a stronger preference to make the soup than she has a preference to refrain from causing her neighbor harm. Because Kay makes soup at t_2, Kay's ignorance, at t_2, does *not* prevent Kay from choosing the option that reflects her preference for making soup over her preference not to cause harm. Therefore, C2 indicates that Kay's ignorance does not have a direct effect on her choice to make soup at t_2. So Kay's ignorance is

not exculpatory for her choice to make soup at t_2, and thus Kay incurs a duty of corrective justice to bear costs in virtue of her choice at t_2.

What this means, then, for states' historical emissions, is that the relevant counterfactuals for applying (ERIE*) to states' historical emissions should describe whether a state's ignorance about the harmful effects of its emissions had a direct effect on the state's policies concerning its production of these emissions. So the relevant counterfactuals are those that describe whether the state's ignorance prevented the state from enacting policies that reflected the state's preferences, rather than describing the effect that the ignorance had on the costs of satisfying its preferences.

The claim about path-dependence—that if states had very early warning of the causes and consequences of climate change, they would have enacted policies to keep their emissions within morally permissible levels—describes the direct effect of states' early ignorance of climate change on their *early* energy policies. If the claim about path dependence is true, we can infer that states have a preference to avoid causing large amounts of global environmental harm. We can also infer that when the costs of satisfying this preference are relatively low, as they were during the early periods of industrialization (we are assuming), the preference would have resulted in policies aimed at avoiding causing large amounts of global environmental harm. Thus, ignorance prevented these states from enacting these policies during the early periods of industrialization. So, if the claim about path dependence is true, then states have not incurred duties of corrective justice in virtue of having produced historical emissions very early on during industrialization.

However, the claim about path dependence describes only the indirect effect that states' early ignorance of the harmful effects of unabated fossil fuel use had on their later energy policies. For example, the United States' early ignorance of the harmful effects of unabated fossil fuel use, by the 1950's, had the effect of raising the costs that the United States must bear to satisfy its preference to avoid causing large amounts of global environmental harm. However, this early ignorance did not prevent the United States in the 1950's from choosing to satisfy its stronger preference to avoid these costs (burdening its citizens with the costs of decarbonization) over its preference to avoid causing large amounts of global environmental harm. Therefore, the claim about path-dependence, even if it is true, does not imply that states have not incurred duties of corrective justice in virtue of having produced historical emissions after industrialization was well under way.

6. Challenge II: Slowness to Change Behavior

Let's turn now to the second challenge to the claim that states' failures to reduce their emissions to morally permissible levels after 1990 means that it is likely that, prior to 1990, states would not have kept their emissions within morally permissible levels, had they known that their emissions would contribute to harmful climate change. Consider that, since 1990, although states have not taken anywhere near sufficient action to address climate change, they have also not done absolutely nothing to address it. Most states have publicly acknowledged that climate change is a problem and that climate change is caused by human activities. Additionally, most states have also signaled their willingness to address climate change by signing international agreements in which they promise to reduce their emissions (including, perhaps most notably, the 2015 Paris Agreement, signed by 195 countries). Moreover, many states have enacted domestic policies designed to curb emissions (e.g., cap and trade in the European Union and the regulation of greenhouse gas emissions under the Clean Air Act in the United States).[31]

Perhaps given the long-term trajectory of states' actions to address climate change, in another twenty or fifty or seventy years, collectively, states will have successfully enacted policies that will have significantly reduced their emissions to levels that will not cause significantly more harmful climate change than what we will be committed to by that time.

But if we take what states have done when they learned that they were contributing to harmful climate change as evidence for what states would have done, prior to 1990, had they known that their emissions would contribute to harmful climate change, then perhaps if states had very early warning of the climate-altering potential of unabated fossil fuel use—in the early stages of industrialization, say—then states would have, within, say, fifty or eighty or one-hundred years, enacted policies that ensured that their emissions remained at levels that would have avoided harmful climate change altogether. If states would have done this, then their historical emissions would have been morally permissible, since the emissions would not have contributed to harmful climate change. Of course, had states learned at some later period that their emissions were contributing to harmful climate change, and if we apply the same fifty or eighty or one-hundred-year period, then perhaps states' historical emissions would have contributed to harmful climate change, and so would have been wrongful.

This creates another puzzle for how to apply (ERIE*) to states' historical

[31] However, since the election of Donald Trump in 2016, both the United States' participation in the Paris Agreement and its regulation of greenhouse gas emissions under the Clean Air Act are now in serious doubt.

emissions. Let's return to the simplified case, *Soup*. Suppose that had Kay had stopped making soup after doing so at t_8, the cumulative effect of the toxins would not have been sufficient to harm V. However, with her choice to make soup at t_9, the cumulative effect of the toxins makes V extremely ill, which costs thousands of dollars to treat.

Suppose that for each time t_1-t_9, had Kay learned just prior to that time that making the soup would risk harming her neighbor, Kay would have made soup five more times before stopping. (Suppose that she needed to wean herself off the very delicious soup.) So suppose that the following counterfactuals are all true:

> CF1: had Kay learned, prior to her first choice to make soup (at t_1), that making the soup would risk harming her neighbor, she would have made soup at t_1-t_5, after which she would have stopped.
>
> CF2: had Kay learned, after at t_1 but prior to t_2, that making the soup would risk harming her neighbor, she would have made soup at t_2-t_6, after which she would have stopped.
>
> . . .
>
> CF5: had Kay learned, after t_4 but prior to t_5, that making the soup would risk harming her neighbor, she would have made soup at t_5-t_9, after which she would have stopped.
>
> . . .

So notice that had Kay learned, any time prior to t_4, that making the soup would risk causing harm, then V would not have been harmed at all. However, had Kay learned at any time after t_4 that making the soup would risk causing harm, then V would have been harmed.

It seems that Kay should *not* incur a duty to bear costs in virtue of her choices prior to t_5, since all the counterfactuals in which Kay would have chosen to make soup at t_1-t_4 are those in which those choices would not have resulted in harm and so would not have been wrongful. It also seems that Kay should incur a duty in virtue of her choice at t_9, since the counterfactuals in which Kay would have chosen to make soup at t_9 are those in which that choice would have resulted in harm, and so would have been wrongful. The question is whether Kay should incur a duty in virtue of her choices t_5-t_8. Had Kay learned, prior to t_4, that her soup-making would

risk causing harm, with respect to each choice t_5-t_8, either Kay would not have chosen to make soup or Kay's choice to make soup would not have been harmful and so would not have been wrongful. Yet had Kay learned, after t_4, that her soup-making would risk causing harm, then Kay would have chosen to make soup at t_5-t_8 and those choices would have contributed to the harm to V, and so would have been wrongful.

I think that Kay should incur a duty to bear costs in virtue of her choices at t_5-t_8. The truth of the counterfactuals in which Kay would have chosen to make soup at t_5-t_8 and in which those choices would have been wrongful creates a relevant asymmetry between Kay (relative to those choices) and V, such that it would be fair for Kay to bear at least some of the cost of rectifying the harm in virtue of those choices. However, given that there are true counterfactuals in which Kay would not have chosen to make soup at t_5-t_8, it is plausible that the cost that Kay incurs a duty to bear on the basis of each of these choices is less extensive than, say, the cost that she has a duty to bear on the basis of her choice at t_9, other things being equal.

Let's return to the circumstances of climate change. The claim that we are considering is that, for each time prior to 1990, had states learned that their emissions would contribute to harmful climate change, they would have continued to emit large quantities of emissions for an additional period of time—say, fifty or eighty or one-hundred years—at which time they would have enacted policies to reduce their emissions considerably. So if states had learned, at some early date (prior to the early part of the twentieth century, say), that their emissions would contribute to harmful climate change, they may have been able to avoid causing any harmful climate change, which would have made the production of their historical emissions morally permissible. On the other hand, if states had learned later (during, say, the second half of the twentieth century) that their emissions would contribute to climate change, they would not have been able to avoid causing climate change, and so their historical emissions would have been morally wrong.

If we assume that the above claims are true, then given my discussion of *Soup,* we can conclude the following. First, states have incurred duties of corrective justice in virtue of having produced historical emissions during the second half of the twentieth century. In addition, the duties of corrective justice that states have incurred in virtue of having produced historical emissions at later dates (e.g., the 1980's) are more extensive than the duties of corrective justice that states have incurred in virtue of having produced historical emissions at earlier dates (e.g., the 1950's), other things being equal. And finally, we can conclude that states have not incurred duties of corrective justice in virtue of having produced historical emissions prior to the early part of the twentieth century.

7. Conclusion

To summarize, I have argued that:

(1) (RIE) is false because (ERIE*) is true. An ignorant causer incurs a duty of corrective justice to bear some of the costs of rectifying a harm or threat of harm that her action causes or contributes to causing when, had the ignorant causer known that her action would cause or contribute to the harm or threat, she would have performed a morally wrong action.

(2) Had states known, prior to 1990, that their emissions would contribute to harmful climate change, it is likely that states would not have kept their emissions within morally permissible levels, given that once states learned, after 1990, that their emissions were contributing to harmful climate change, states did not reduce their emissions to morally permissible levels.

(3) Therefore, it is likely that states have incurred duties of corrective justice to bear costs in virtue of having produced historical emissions.

(4) However, given both (a) the path-dependence of fossil fuel use and (b) the possibility that meaningful action to avoid harmful climate change would have been delayed (rather than completely nonexistent), states may not have incurred duties of corrective justice in virtue of having produced historical emissions prior to the early part of the twentieth century.

References

Aristotle. (1999). *Nicomachean Ethics* (T. Irwin, Trans). Indianapolis, IN: Hackett.

Caney, S. (2005). Cosmopolitan justice, responsibility, and global climate change. *Leiden Journal of International Law*, 18(4), 747–775.

------ (2012). Climate change and the duties of the advantaged, *Critical Review of International Social and Political Philosophy*, 13(1), 203–228.

Coleman, J.L. (1995). The practice of corrective justice. In D. Owen (Ed.), *Philosophical Foundations of Tort Law* (pp. 53-72). Oxford: Clarendon Press.

-----(2001). Tort law and tort theory. In G. Potsema (Ed.), *Philosophy and the Law of Torts* (pp. 183-21). Cambridge: Cambridge University Press.

Cripps, E. (2013). *Climate change and the moral agent: Individual duties in an interdependent world*. Oxford: Oxford University Press.

Gardiner, S. (2011). *A perfect moral storm.* Oxford: Oxford University Press.

McMahan, J. (1994). Self-defense and the problem of the innocent attacker. *Ethics,* 104(2), 252–290.

McMahan, J. (2002). *The ethics of killing: Problems at the margins of life*. Oxford: Oxford University Press.

Parfit, D. (2011). *On what matters, vol. 1.* Oxford*:* Oxford University Press.

Peels, R. (2014). What kind of ignorance excuses? Two neglected issues. *The Philosophical Quarterly*, 64(256), 478–496.

Risse, M (2008). Who should shoulder the burden? Global climate change and common ownership of the earth. *Harvard Kennedy School of Government Faculty Research Working Paper Series*.

Scanlon, T.M. (1998). *What we owe to each other.* Cambridge, MA: Harvard University Press.

Schüssler, R. (2011). Climate justice: A question of historic responsibility? *Journal of Global Ethics,* 7(3), 261–78.

Shue, H. (1999). Global environment and international inequality. *International Affairs*, 75(3), 531–545.

Singer, P. (1972). Famine, affluence and morality. *Philosophy and Public Affairs*, 1(3), 229–243.

Stilz, A. (2011). Collective responsibility and the state. *The Journal of Political Philosophy*, (19)2, pp. 190–208.

Tadros, V. (2011). *The ends of harm*. Oxford: Oxford University Press.

Thomson, J.J. (1990). *The realm of rights*. Cambridge: Harvard University Press

Vanderheiden, S. (2008). *Atmospheric justice: A political theory of climate change*. Oxford: Oxford University Press.

WRI. (2014). *Climate analysis indicators tool: WRI's climate data explorer*. World Resources Institute. Retrieved from http://cait2.wri.org.

Wündisch, J. (2016). Does excusable ignorance absolve of liability for costs? *Philosophical Studies,* doi:10.1007/s11098-016-0708-1.

Zellentin, A. (2015). Compensation for historical emissions and excusable ignorance. *Journal of Applied Philosophy,* 32(3), 258–274.

Zimmerman, M. J. (1997). Moral responsibility and ignorance. *Ethics*, 107 (3), 410–426.

Katie Steele[1]

The distinct moral importance of acting together

This review essay engages with Garrett Cullity's argument that there is a fundamental moral norm of cooperation, as articulated in *Concern, Respect, & Cooperation* (2018). That is to say that there is moral reason to participate in collective endeavours that cannot be reduced to other moral reasons like promoting welfare. If this is plausible, all the better for solving collective action dilemmas like climate change. But how should we understand a reason of participation? I supplement Cullity's own account by appealing to the notion of 'team reasoning' in game theory. Even if not an adequate notion of rationality, adopting the team stance—deriving individual reason to act from what a group may together achieve—may well have distinct moral importance.

[1] School of Philosophy, Australian National University, katie.steele@anu.edu.au

1. A plurality of moral foundations

In *Concern, Respect, & Cooperation*, Garrett Cullity defends a pluralist account of morality, whereby moral reasons for behaviour and attitudes rest on more than one foundation, none of which is reducible to others. Two of the pillars on which he builds his account are commonly taken to be in tension: concern for others' welfare, and respect for their agency. Controversially, Cullity sees a role for both. But more intriguing again is the third pillar, which he presents most simply (i.e., minus the caveats) as follows:

> "Our worthwhile collective action calls for my action of joining in."

In Cullity's words (p. 52): "I want to take seriously the idea that when people manifest this form of decency, they are following norms that are just as fundamental to morality as the norms for concern and respect. The question 'Why join in worthwhile collective actions?' is like 'Why help people who need it?' or 'Why allow others to live their own lives?'"

But these questions are not obviously on a par. The response "because they are persons" seems adequate for the latter two questions. But it is not an evidently adequate response to the first. What is so important, after all, about acting together? It is not obvious that a fundamental way of recognising others as persons is to team up with them to pursue joint ends.

As with many aspects of this rich book, whether or not one agrees with Cullity's inclusion of cooperation as a moral foundation, the idea is original and fruitful. It is worth considering what a fundamental norm of cooperation would plausibly look like. In this comment I initially lay out Cullity's own characterisation, and ultimately build from this a much fuller picture of the norm, albeit pushing it in a direction that Cullity may not endorse. I claim, however, that there is no alternative. The more salient ways of understanding the norm are neither convincing nor helpful for filling in crucial details. My own proposal draws on the rich findings of game theory regarding the tragedies of individualist reasoning, or what is known as *collective action problems*.

2. When there is reason to join in

The fundamental moral norms, on Cullity's view, furnish reasons for action that are by definition non-derivative, and yet not necessarily trumping. They are merely *pro tanto* reasons: *some* consideration in favour of doing *X* rather than *Y* that may nonetheless be defeated by other moral or nonmoral considerations. Reasons of this

sort can be rather blunt. For instance, there is plausibly a reason to help any and all strangers in severe need; where it gets complicated is how this reason is weighed against other competing reasons to act otherwise. That said, when it comes to cooperation, even *pro tanto* reasons surely require qualification. Unlike helping strangers, the idea that there is even *some* consideration of decency to join in any and all collective efforts is simply not compelling.

Cullity's norm of cooperation is indeed a qualified one; the nuance lies in his specification of *worthwhile* collective actions. To begin with, whether a collective action is worthwhile depends on other moral and non-moral reasons. This clearly rules out collective acts of aimless harm and wanton destruction. There is no reason for any agent, whether an individual or a group, to pursue such ends. But Cullity goes further in suggesting that whether a collective action is worthwhile may be sensitive to *who* is concerned and what other reasons bear on their actions (p. 55). For instance, it may be worthwhile for *me* to join a choir, because the practice hall is merely five minutes from my house, but not worthwhile for *you* to join the choir, despite your similar love of singing, because the practice hall is too long a commute for you and the outing would thus consume too many resources.

The single term "worthwhile" thus plays an important and complex role in determining who, if anyone, has reason to act together to pursue some end. In short, those who are enjoined to participate in a collective action are those who *similarly have reason* to pursue the joint endeavour; they are of common "kind *K*" with respect to the endeavour, to use Cullity's words (p. 55). Cullity gives the following examples:

> "... where the action is one of collective self-interest in producing a public good, the group may be constituted by people for whom the benefit being produced by the collective action outweighs the cost of contribution. Where the action is one of group beneficence, the group may be constituted by people with the capacity to contribute without serious personal cost..."

These examples further suggest that an individual's own reasons for participating in a collective action are closely bound up with those of others. For instance, whether or not the benefits of producing a public good outweigh the costs for any given individual may depend on how many people are similarly placed, since this affects how much of the public good would stand to be produced and thus the size of the benefit that is weighed against the personal cost.

These details go a long way towards identifying those collective actions that call for joining in and are suggestive of a fundamental norm of cooperation. But there remain some critical ambiguities. In particular, it is not clear what, exactly, are the criteria for individuals being similarly placed to achieve something together. For

instance, does being similarly placed mean that individuals have identical options for how to act as well as identical sets of reasons for pursuing those options? It seems not, since, in the spirit of *pro tanto* reasons, Cullity allows that you and I may be similarly placed even if you have reason to join just one collective action, *C,* whereas I have reason to join *C* in addition to some further collective actions (see, e.g., p. 54). In that case our sets of reasons are not identical. To what extent then, must our options and reasons coincide for us to count as similarly placed or of common kind *K*? We are owed a response to this question, since the notion plays a pivotal role in defining worthwhile collective action. This is not just a matter of stipulation; what is needed is a bigger-picture story for when and why a norm of cooperation is integral to our moral lives.

3. In search of a bigger picture

One might worry that no bigger-picture story can nor need be given for a fundamental moral norm. After all, such norms, by definition, cannot be derived from other norms. To motivate a fundamental norm, it seems wiser to seek paradigm cases of the norm's manifestation, that is, cases where the norm in question, and no other, clearly furnishes a reason for action of moral significance. For instance, the norm of concern can be motivated by cases where an agent has the opportunity to greatly relieve suffering. Simply in contemplating such cases, we appreciate that relieving suffering provides reason to act, and this is a matter of living decently.

The problem is that there are not such obvious paradigm cases for a norm of cooperation. Such cases rather require careful construction; hence the need for a bigger-picture story. To be sure, there are vivid cases of worthwhile collective action as defined above. The difficulty is in identifying worthwhile collective actions for which, *but for* a norm of cooperation, one cannot explain what seems a compelling reason to join in.

Consider, for instance, the following case.

> *Two Hikers*: Two friends are hiking in the mountains when they come across a person trapped under a boulder. The closest friend immediately tries to lift the boulder, but its weight proves too much for one person alone. The other friend sees that her participation will allow the boulder to be lifted. She thus has ample reason to join in.

While clearly a case of worthwhile collective action, *Two Hikers* does not offer clear support for there being a fundamental norm of cooperation. The second friend finds herself in the position of being able to either save someone's life by joining in the

effort to lift the boulder, or else continue unfatigued in exploring the mountains. The former action enhances wellbeing much more than the latter. So, concern for wellbeing clearly favours joining in the lifting. There is no need to appeal to a reason of cooperation to explain why the friend has strong reason to join in.

To motivate the norm of cooperation, Cullity himself appeals primarily to cases like the following, where the agent in question would make next to no difference in joining in:

> *Many Hikers*. Numerous friends are hiking in the mountains when they come across a person trapped under a boulder. The closest two friends immediately start lifting the boulder and will clearly prevail. Moreover, the subsequent contributions of the other friends would be negligible in terms of relieving the burden of the first-movers. Nonetheless, all have reason to join in.

In this case there is very limited reason of concern to join in, since the trapped person would be saved regardless, and the burden of the first-movers would be reduced only marginally. There is no obvious reason of respect to join in either. Indeed, at the limit where joining in makes no difference whatsoever, there cannot be any reason of either concern or respect to do so.

The *Many Hikers* case is therefore of the right sort to support a fundamental norm of cooperation and reveal its nature. The problem, however, is that it is not clear that there is *any* reason for the latecomers to join in the boulder lifting. Is it really worthwhile to pile on here? Not obviously so, at least. To be sure that this is a paradigm case for the norm of cooperation, we need a bigger-picture story that illuminates why *Many Hikers* has the right features to make joining in both desirable and not attributable to other sorts of reasons.

Is there such a story in support of *Many Hikers*? In the remainder of this section, we will pursue a couple of salient possibilities that are ultimately unsuccessful. The first candidate is one that appeals to the unique fellowship that comes from acting together. This fellowship, in and of itself, regardless of any difference the agent makes to the group outcome or the burdens of other group members, is valuable and is thus reason to join in. Or so the story might go. Now one might object that even if *Many Hikers* isolates a reason to do with fellowship for joining in, it is only a very weak reason; the fellowship story is not convincing insofar as there is *strong* reason for the friends to join in lifting the boulder. These friends are out taking a hike together, after all, so presumably they already have fellowship aplenty. But let us grant that fellowship is here a strong moral reason to join in. The bigger objection is that the fellowship associated with a joint endeavour should already be counted amongst its wellbeing outcomes, even if typically overlooked because its contri-

bution to wellbeing goes unnoticed. It is thus covered by the norm of concern. That is, we need not appeal to some further norm of cooperation to accommodate strong reasons of fellowship. On this count, *Many Hikers* is, if anything, a paradigm case for a less familiar aspect of the norm of concern.

The second candidate story appeals rather to the expression of equality that comes with joining in. Cullity proposes something along these lines. He emphasises that agents who are similarly placed to achieve some collective outcome have reason to act together, simply *because they are similarly placed*. In *Many Hikers*, for instance, even though the boulder can be lifted by just two persons and any further contribution by others would make negligible difference, since all the friends are similarly placed with respect to this outcome, they all have reason to express their non-exceptionalism, as it were, by joining in. While there is something to this story, as it stands it does not help in articulating the details of the norm of cooperation. That is, we cannot hope to refine what it means for individuals to be "similarly placed" by appeal to a story that emphasises what equality demands of those who are similarly placed. Moreover, one might wonder why such a norm of equality is needed in the first place. Agents who take reasons of concern and respect adequately into account already seem to treat all moral subjects as equals. The norm of concern asserts that the wellbeing of *all* is similarly worthy of promotion. And the norm of respect asserts that *all* similarly deserve not to have their self-expression interfered with. Do we need to postulate a further demand of equality? Perhaps, but more must be said for this to be convincing.

4. Circumventing the tragedies of individual reasoning

It helps to focus on how a norm of cooperation differs from those of concern and respect. The three norms may each be fundamental, but they need not all have the same target or operate in the same way. The norms of concern and respect are similar in that they are about what are better and worse ways the world might be. The world is better when a person has more wellbeing or when her self-expression is not interfered with, all else being equal. These norms thus bear directly on the outcomes of action. In *Two Hikers*, for instance, both friends have reason to join in lifting the boulder because, on the understanding that the other will also do her part, they will each make a difference to the outcome. The trapped person will thereby be released—a big boost in her wellbeing. But the norm of cooperation is different. It is not about the goodness, even broadly construed, of action outcomes. It does not furnish first-order reasons for action in this way. It is best conceived, rather, as

furnishing second-order reasons for action. When agents perceive that they have similar reasons for joining in a collective action, on the norm of cooperation these reasons themselves generate a further reason to join in.

The most plausible way to develop this story is by appeal to familiar lessons from game theory regarding the tragedies of individualist reasoning. The tragic cases, known as *collective action problems*, are ones where individuals together choose a Pareto-inferior act (one that is worse for someone and not better for anyone than some alternative) due to the limitations of each reasoning unilaterally about what to do. These problems come in different forms. There is the well-known *Prisoners' Dilemma*, a particularly troubling scenario whereby it is rational for each individual, reasoning unilaterally, to defect from some group action, regardless of what others do, even though it would be better for each if all joined in the group action. Other sorts of collective action problems arise not from the incentive to knowingly "free ride" in this way but rather from uncertainty about what others will do. The latter are referred to as *Coordination Dilemmas*. In *Many Hikers*, for instance, it is merely fortuitous that there are two first-movers who lift the boulder. While any of the friends would presumably be willing to help lift the boulder if they knew that they would be a difference-maker, none has this knowledge. It may turn out that all think it so unlikely that they will be the difference-maker that, tragically, none has sufficient reason to join in.

It would seem then that it is this feature of cases—their being collective action problems—that makes them candidates for motivating a fundamental norm of cooperation.[2] Joining in seems genuinely worthwhile, since, from a shared group perspective, all have sufficient first-order reason to pursue some joint endeavour when others do too. From each isolated individuals' perspective, however, the first-order reasons do not suffice for joining in. *But for* a reason of cooperation, there is thus insufficient reason for an individual to join in. This further second-order reason, as it were, is needed for individuals to do what is optimal from their shared group perspective.

Note that this story is not quite so compelling when it comes to coordination dilemmas as compared to free-rider prisoners' dilemmas. That is because coordination dilemmas admit of more sophisticated resolutions. After all, if all join in some collective endeavour that only requires a small number to get the job done, then time and labour are wasted. Still, better this excess than the job not being done

[2] Indeed, Cullity's own illustrative cases for the norm of cooperation have the form of collection action problems even if he does not present them as such. His trapped-under-rubble case (p. 222) is analogous to *Many Hikers* and can thus be conceived as a *coordination dilemma*. His consumer boycott case (p. 231) and the case of action against climate change (p. 233) are presented, roughly speaking, as *prisoners' dilemmas*.

at all. But if the individuals were able to communicate in a timely fashion, they might be able to draw straws, say, to determine who should join in. In *Many Hikers*, while there is no communication, two friends happen to choose to start lifting the boulder; here it seems perverse that the remaining friends still have a reason of cooperation to join in when it is clear they will make no difference. But one could argue that this reasoning serves as insurance in that it ensures the boulder is lifted, were an insufficient number to take it upon themselves to join in, as could well have happened.[3] In prisoners' dilemma cases, by contrast, those who join in due to a reason of cooperation always make a difference. For instance, reducing one's personal carbon emissions *does* make a difference, even if not enough of a difference, but for a reason of cooperation, to outweigh the costs of this personal sacrifice. So, prisoners' dilemmas make for clearer-cut paradigm cases for the norm of cooperation.

5. A norm of morality or rationality?

The appeal to game theory may help in refining a fundamental moral norm of cooperation, but it raises the question: Is this a *moral* norm or merely one of rationality? The norm we have described is reminiscent of a revisionist notion of rational choice in collective settings—known as *team reasoning*—that was first defended by Robert Sugden (1993). The idea is that rational individuals, even non-altruistic ones, would not consider what is best for themselves in isolation but would rather consider what part they should play in a group effort to produce outcomes that are best for all. In this way, players would overcome the prisoners' dilemma. A related proposal defended by Lawrence Davis (1977), also intended to overcome the prisoners' dilemma, is that rational individuals choose that which is best for all rational individuals who face the same options to choose.

The vast majority of game theorists, however, are not convinced by either of these augmented notions of rational choice. Binmore (1994) articulates the widespread view that these proposals depart from the minimal notion of rational choice by introducing substantial assumptions about what agents should value or else believe about other players' choices. But that is precisely why these proposals may be better seen as animating *moral* norms that provide individuals with further substantive reasons for choice. The latter proposal which invokes what we might dub *symmetric reasoning* seems particularly promising in this regard. Agents have reason to act as if others similarly placed will act similarly, not because the evidence already suggests that this is indeed how others will act but rather by way of *creating*

[3] Note that this marks a point of divergence with Cullity, who does not to think it problematic or worthy of explanation that individuals have a reason to join in a collective action, similarly placed to others as they may be, when it is clear that they will make no difference whatsoever.

evidence that this is how others will act. After all, for a nuanced moral theory like Cullity's, even in the best circumstances consisting of perfectly moral agents with full information, there would otherwise be many tragic collective circumstances.

References

Binmore, K. (1994). Playing Fair: Game Theory and the Social Contract 1, Cambridge, MA: MIT Press.

Cullity, G. (2018). Concern, Respect, & Cooperation. Oxford: OUP.

Davis, L. (1977). Prisoners, Paradox and Rationality. American Philosophical Quarterly, 14: 319–327.

Sugden, R. (1993). Thinking as a Team: Towards an Explanation of Nonselfish Behaviour. Social Philosophy and Policy 10: 69–89.

Gustaf Arrhenius,[1] Mark Budolfson[2] & Dean Spears[3]

Does Climate Change Policy Depend Importantly on Population Ethics? Deflationary Responses to the Challenges of Population Ethics for Public Policy

Choosing a policy response to climate change seems to demand a population axiology. A formal literature involving impossibility theorems has demonstrated that all possible approaches to population axiology have one or more seemingly counterintuitive implications. This leads to the worry that because axiology is so theoretically unresolved as to permit a wide range of reasonable disagreement, our ignorance implies serious practical ignorance about what climate policies to pursue. We offer two deflationary responses to this worry. First, it may be that given the actual facts of climate change, all axiologies agree on a particular policy response. In this case, there would be a clear dominance conclusion, and the puzzles of axiology would be practically irrelevant (albeit still theoretically challenging). Second, despite the impossibility results, we prove the

[1] Institute for Futures Studies & Department of Philosophy, Stockholm University, gustaf.arrhenius@iffs.se.
[2] University of Vermont Gund Institute for Environment & Department of Philosophy and Australian National University, mark.Budolfson@uvm.edu
[3] Department of Economics and Population Research Center, University of Texas at Austin. Indian Statistical Institute, Delhi Centre. IZA Institute of Labor Economics. Institute for Futures Studies, dspears@utexas.edu.

possibility of axiologies that satisfy bounded versions of all of the desiderata from the population axiology literature, which may be all that is needed for policy evaluation.

*

> "To plan an appropriate response to climate change, it is important to evaluate each of the alternative responses that are available. How can we take into account changes in the world's population? Should society aim to promote the total of people's wellbeing in the world, or their average wellbeing, or something else? The answer to this question will make a great difference to the conclusions we reach."
> (Pachauri, Mayer, & Intergovernmental Panel on Climate Change (2015)).

1. Introduction

The International Panel on Climate Change (IPCC) and some leading philosophers and economists have expressed unease about the implications of population change for evaluating responses to climate change and other intergenerational policy challenges. Their unease derives from a common view among those who investigate the questions of population ethics, that is, theories about the value of outcomes where the number of people, the quality of their lives, and their identities may vary. The view is that we do not know what to do about intergenerational policy until we know what to do about population ethics. John Broome, in particular, has prominently voiced the concern that climate policy could turn critically on unresolved questions in population ethics.[4] The worry expressed by Broome and reflected in the quote from the IPCC above might be stated as follows:

> **Worry:** Because climate change, climate policy, the size of the population, and population policy all may have effects on one another, and because population ethics is so theoretically unresolved as to permit a wide range of reasonable disagreement about social evaluation, our ignorance of the correct population ethic implies serious practical ignorance about what climate policies to pursue.[5]

[4] See, e.g., Broome (1992), (2004), ch. 1, and (2012b).
[5] "We do not know what value to set on changes in the world's population. If the population shrinks as a result of climate change, we do not know how to evaluate that change. Yet we have reason to think that changes in population may be one of the most morally significant effects of climate change. The small chance of catastrophe may be a major component in the expected value of harm caused by climate

In this chapter, we argue that the Worry is not obviously well-founded: we may already know enough to make good choices about climate policy even without further progress in population ethics, and further progress might not make much difference to the conclusions that are ultimately correct. More generally, we highlight some reasons – some philosophical, some empirical – why intergenerational policymaking might not be very sensitive to classic arguments from population ethics in the way that have often been assumed.

To understand why the IPCC and many others share the Worry, we must begin by noting that intergenerational policymaking seems to require a concept of goodness that aggregates consequences for many different people (perhaps even non-humans), with different properties, living at different times. Most of these people are not yet alive. Most of them will only ever be born depending on which particular climate policy is chosen. But any response to climate change requires integrating over the consequences for all of them.

For example, consider the Integrated Assessment Models (IAMs) of climate policy constructed by economists and other researchers. In 2018, William Nordhaus was awarded the Economics prize to the memory of Alfred Nobel, partly for his family of climate policy IAMs. IAMs like Nordhaus' choose an optimal carbon tax policy, balancing the disadvantages of more expensive energy with the advantages of reduced global warming. More broadly, reducing fossil fuel consumption could increase present-day economic costs for both poor people and rich people; could slow economic growth and poverty alleviation in the developing world; and could prevent future harm from temperature increases — increases which will help some people, but hurt many more people, and have consequences for inequality. The socially optimal carbon tax or fossil fuel policy depends on taking all of these and other relevant factors into proper account – which seems to require weighing the aggregate of these consequences conditional on different policy options.

So, choosing a policy response to climate change seems to demand an aggregative concept of goodness — an axiology. Those who study axiology have devoted considerable theoretical attention to population ethics: to the questions of how rankings of aggregate social goodness extend to ranking outcomes in which different people and different numbers of people exist. Parfit (1984) identified many of the core questions of population ethics, which are widely regarded to remain open. A number of candidate resolutions have been offered in the literature, but a formal literature involving impossibility theorems — led by Arrhenius (2000a), (2000b) and sub-

change, and the loss of population may be a major component of the badness of catastrophe. ... So we face a particularly intractable problem of uncertainty, which prevents us from working out what we should do. Yet we have to act; climate change will not wait while we sort ourselves out" (Broome (2012a), pg. 183-185).

sequent work — has demonstrated that each approach (and all possible approaches) has one or more seemingly counterintuitive implication. These theorems appear to show that our considered moral beliefs are mutually inconsistent, that is, that necessarily at least one of our considered moral beliefs is false. Since consistency is, arguably, a necessary condition for moral justification, it may appear that we are forced to conclude that there is no moral theory which can be justified. Moreover, we would then lack the theoretical tools needed to evaluate climate options in which the number of people, the quality of their lives, and their identities will differ.

In Section 2 we introduce in more detail these paradoxes and the related population axiology literature, with special focus on Parfit's well-known Repugnant Conclusion. With this introduction in hand, Section 3 offers the first and simplest of two deflationary responses to the Worry: it may be, given the actual facts of climate change, that all axiologies agree on a particular policy response. In this case, there would be a clear dominance conclusion, and the puzzles of population ethics would be practically irrelevant (albeit still theoretically challenging). Section 4 offers the second more complex deflationary response: despite the impossibility results from Arrhenius, it is nonetheless possible to prove the possibility of axiologies that satisfy *bounded* versions of all of the desiderata from the population ethics literature that Arrhenius's proofs marshal. In this way, an incomplete population axiology that is defined over the practically relevant bounded space can avoid the Repugnant Conclusion and satisfy other relevant bounded versions of the adequacy conditions in population ethics. Assuming that we only need to consider the bounded versions of the adequacy conditions when we consider policy issues, and that analogous impossibility theorems cannot be proved in the bounded domain, we can for practical purposes put the impossibility theorems that have haunted population ethics to the side.

These deflationary responses do not show that theoretical progress towards population axiology should not continue. Indeed, as we shall show below, an important consequence of the second deflationary response is that it shows the need of more scrutiny of what the core intuitions behind the adequacy conditions in population ethics really are, and further investigation of axiologies on bounded domains. The upshot of this paper is that responding to climate change, and policy analysis more generally, may not need to wait for greater consensus in population ethics on unbounded domains, and that the possibility of deflationary responses to the impossibility theorems deserves further attention.

3. Population axiology and the Repugnant Conclusion

Population axiology concerns how to evaluate populations of different sizes in regard to their goodness: how to assign a value to increases and decreases in population size. The first few papers in this field were not published until the late 1960s and it did not become a significant field until Derek Parfit's famous book *Reasons and Persons*, published in 1984. It is now a very lively field of inquiry.

As John Broome has noted, policymakers seem to almost universally ignore the effects of policy on population size. Why do they ignore it? One possible explanation is that many people have what Broome calls the **Intuition of Neutrality**, which holds that adding a person to the world's population makes the world neither better nor worse.[6] Hence, effects on population size is something that we do not need to think about, or if we do need to think about it, it is because it makes people's lives better or worse; other than that, having a bigger or smaller population does not make any difference to the value of outcomes.

There are likely to be limits to Neutrality. For example, most people would probably agree that if population growth leads to having many people with very bad lives, then that would make the world worse. In light of this, we think that among those people who have intuitions in this neighbourhood, it is more likely that they endorse the more limited **Asymmetry Intuition** (which also appeared earlier in the literature):[7] We have no moral reasons for or against creating people with positive welfare stemming from the welfare these people would enjoy, but, on the other hand, we have moral reasons against creating people with negative welfare stemming from the negative welfare these people would suffer. Hence, those people are neutral only about adding people with positive welfare.[8] However, assuming that future people have positive or neutral welfare, the idea is that population size is neutral in terms of value and that we can ignore this aspect when considering different policies.

- Population *B* consists of a number of people with very low positive welfare, and

[6] For a more detailed discussion of the neutrality intuition, see Broome (2004), (2010).
[7] How many people in fact endorse the Asymmetry is an empirical question; in one recent survey Spears (2019) finds that only a minority of respondents do. The study also provide suggestive evidence for weaker versions of the Asymmetry focused on the weight of suffering and parental procreative autonomy, as discussed in Arrhenius (forthcoming), section 9.5.
[8] This formulation is from Arrhenius (forthcoming), (2000b). For earlier formulations, see McMahan (1981); Parfit (1982).

- Population *C* is a population of the same size as *B* but made up of people with very high welfare.

According to Neutrality and Asymmetry, either adding *B* or adding *C* to *A* each would make the resulting populations equally good, given full comparability.[9] But surely, when other things are equal, it must be better to create people with very high welfare rather than people with very low welfare. Hence, population *A+C* is better than population *A+B,* which contradicts Neutrality and Asymmetry. So they are false. And because they are false, climate policy-making must consider population size in its evaluation of outcomes.

The opening quotation from the IPCC listed two alternative approaches to aggregating welfare. One approach is Total Utilitarianism: when we evaluate future populations in respect of population change, we look at the total welfare in the different possible outcomes and rank them by how much total welfare they contain. According to this view, we should maximize the total amount of welfare in the world. So if there are more people with lives worth living, then that is better.

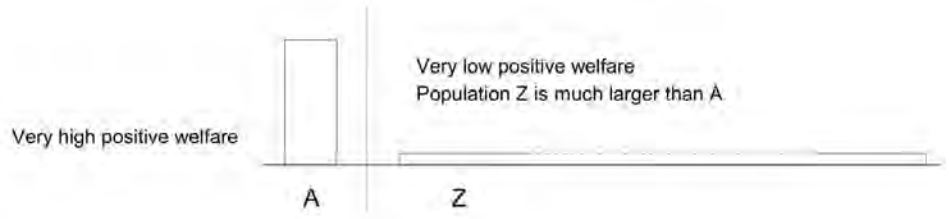
Now a problem with this view is that it has a number of very counterintuitive implications. Much theoretical attention in population ethics has focused on a particular implication of Total Utilitarianism. Total welfare can be increased in two ways when the size of the population is no longer fixed: by keeping the population at a constant size and making people's lives better, or by increasing the size of the population by adding new people with lives worth living. So, according to Total Utilitarianism, a future with an enormous population with lives barely worth living could be better than a future with a smaller population with very high individual quality of life. But the idea that it would be better to radically increase the world's population at the expense of future people's individual welfare seems repugnant to many, and rather a reason to reject Total Utilitarianism. It is an instance of Parfit's infamous Repugnant Conclusion:

> **Repugnant Conclusion:** For any population consisting of people with very high positive welfare, there is a better population in which everyone has a very low positive welfare, other things being equal.[10]

[9] Giving up full comparability isn't sufficient to save the neutrality and asymmetry intuition, see Arrhenius (forthcoming) and Broome (2004).

[10] Here's how Parfit (1984), p. 388 formulates the conclusion: "For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better, even though its members have lives that are barely worth living." Hence, our formulation from Arrhenius (2000b) is more general than his. The

Figure 1. The Repugnant Conclusion



In Figure 1, the width of each block represents the number of people; the height represents their lifetime welfare. Dashes indicate that the block in question should be much wider than shown, that is, the population size is much larger than shown. These populations could consist of all the past, present and future lives, or all the present and future lives, or all the lives during some shorter time span in the future such as the next generation, or all the lives that are causally affected by, or consequences of a certain action or series of actions, and so forth.

All the lives in the diagram have positive welfare, or, as we also could put it, all the people have lives worth living. The A-people have very high welfare whereas the Z-people have very low positive welfare. The reason for this could be that in the Z-lives there are, to paraphrase Parfit, only enough ecstasies to just outweigh the agonies, or that the good things in those lives are of uniformly poor quality, *e.g.*, eating potatoes and listening to Muzak.[11] Or it could be that the Z-people have quite short lives as compared to the A-people. We could imagine that in A, the people live for, say, 80 years whereas in Z the average life expectancy is, say, 40 years, like in some developing countries in the 1970s. However, because there are many more people in Z, the total sum of welfare in Z is greater than in A. Hence, a theory like Total Utilitarianism, according to which we should maximize the welfare in the world, ranks Z as better than A --- an instance of the Repugnant Conclusion.

As the name indicates, many people find the Repugnant Conclusion a reason to reject Total Utilitarianism; to these, the idea that we can make the world better by expanding the population at the expense of future people's individual quality of life seems very counterintuitive. The Repugnant Conclusion has sometimes been taken

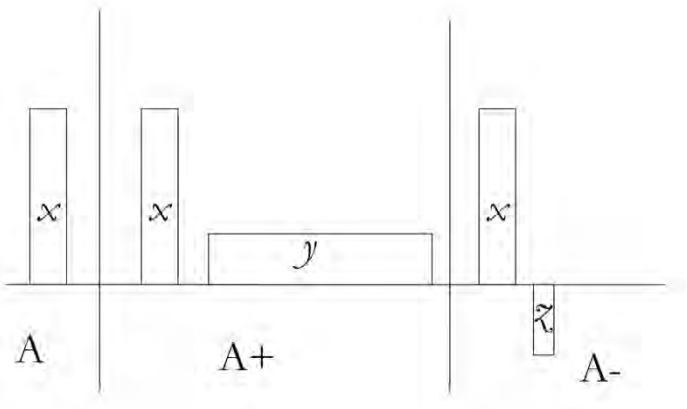
ceteris paribus clause in the formulation is meant to imply that the compared populations are roughly equal in all other putatively axiologically relevant aspect apart from individual welfare levels. Although it is through Parfit's writings that this implication of Total Utilitarianism has become widely discussed, it was already noted by Henry Sidgwick (1907), p. 415, before the turn of the century. For other early sources of the Repugnant Conclusion, see Broad (1979), pp. 249–250, McTaggart (1927), pp. 452–453, and Narveson (1967).

[11] See Parfit (1984), p. 388 and Parfit (1986), p. 148.

in the literature as the major objection to Total Utilitarianism that allegedly disqualifies it as a plausible axiology.[12]

The other approach mentioned by the IPCC is to maximize *average* welfare in the world. This is what Average Utilitarianism tells us to do. Returning to Figure 1, in the case of the A and Z populations the average principle recommends A, because average welfare is much higher in A than in Z. Hence, Average Utilitarianism avoids Parfit's Repugnant Conclusion, which may seem to count in its favour.[13] Unfortunately, it has even worse problems. One problem with maximizing average welfare is that it implies that it can be better to add one group of people to the population rather than some other group, even if each person in the former group has a life that is not worth living and each person in the latter group has a life that is worth living. This is illustrated in Figure 2:

Figure 2. The Sadistic Conclusion



[12] There are other implications of Total Utilitarianism in population ethics that arguably are even more counterintuitive than the Repugnant Conclusion, see e.g., Arrhenius (forthcoming), (2000b), (2011). More on this below.

[13] As explained below, Budolfson & Spears (2018c) have argued that Parfit's initial illustration is only a subset of the classical Repugnant Conclusion, and that we should understand it to include a version (based on addition to a base population, explained in their paper) that is implied by Average Utilitarianism and other axiologies that are commonly taken to avoid the repugnant conclusion. Throughout this section, for clarity we maintain the standard terminology in the population literature, except where it is clear we are discussing the argument of Budolfson and Spears. Anglin (1977) and Arrhenius (2000b), ch. 3, 10 note that Average Utilitarianism implies a version of the Repugnant Conclusion to the effect that that for any population with very high welfare, it can be worse to add this population rather than a population with very low welfare. As Anglin summarized simply: "in some cases the average principle also leads to the Repugnant Conclusion" (p. 746).

Here, we have the A population where the x-people's quality of life is very high. Assume that we can either increase population either by adding the y-people that have quite low but positive welfare—their lives are worth living—or by adding the z people, all of whom are suffering horribly—their lives are not worth living.

Because adding a lot of people with very low but positive welfare can decrease the average welfare of the population more than adding fewer people suffering horribly, it might be better, according to Average Utilitarianism, to add the suffering lives (the z-people) rather than the lives worth living (the y-people). Again, we have a very counterintuitive conclusion on our hands. This is what Arrhenius called the Sadistic Conclusion:

> **Sadistic Conclusion:** It can be better to expand the population by adding people with negative welfare rather than adding people with positive welfare, other things being equal.[14]

The path away from the Repugnant Conclusion towards the Sadistic Conclusion illustrates the puzzles that motivate the Worry. There may be no principle for evaluating populations that is not in some way very counterintuitive. This possibility was originally raised informally by Parfit, who presented a number of paradoxes in population ethics. Much of the important theoretical progress since then has been in formalization of these conclusions and axiologies, as well as many others, and their integration into rigorous proofs.

This literature has progressed, at first, through a dialogue in which researchers proposed and formalized alternative population axiologies (Greaves (2017)). Each was specially formulated to avoid versions of the Repugnant Conclusion, and then further explored by researchers. So, Ng (1989) introduced a variable-value axiology, in which the average utility of a population is inflated by a positively increasing, concave function of population size, such that social evaluation asymptotes from nearly-Total Utilitarianism to nearly-Average Utilitarianism as population size increases. Like Average Utilitarianism, Ng's theory does not escape the Sadistic Conclusion. Blackorby & Donaldson (1984) and later Blackorby, Bossert, & Donaldson (1995) propose Critical-Level Generalized Utilitarianism; this approach also avoids the Repugnant Conclusion at the cost of implying the Sadistic Conclusion. Other approaches, such as Sider (1991)'s theoretical example of Geometrism, or Asheim & Zuber (2014)'s Rank-Dependent Generalized Utilitarianism, attend to people's *rank* within a population, like maximin does. These avoid the Repugnant Conclusion, but have other implausible properties, including in

[14] See e.g., Arrhenius (2000b), (2000a).

cases where population size does not change, such as recommending redistribution from the worst off to the best off in some cases.[15]

None of these proposals has resolved the paradoxes. Led by Arrhenius (2000b), the literature has now established a number of impossibility theorems that demonstrate that no axiology can simultaneously satisfy various sets of very compelling adequacy conditions or principles. Trying to satisfy all of them at the same time leads to contradiction. These conditions are of the type that we have been considering—for example, what Arrhenius calls the Egalitarian Dominance Condition, which states that one population A is better than another same-sized population B if A is perfectly equal and every person in A is better off than every person in B. This condition is incompatible with several other compelling conditions, including conditions that are formulated to rule out the Repugnant and the Sadistic Conclusions. The first and perhaps most well-known of these impossibility theorems is the following:

> **Impossibility Theorem** (Arrhenius (2000a)): There is no welfarist axiology that satisfies the Dominance, the Addition, and the Minimal Non-Extreme Priority Principle and avoids the Repugnant, the Sadistic and the Anti-Egalitarian Conclusion.[16]

Although we refer the reader to the formal statement by Arrhenius (2000a), we emphasize here that each of the conditions listed in the theorem is intuitively compelling. For example, the Dominance Condition is simply that if everyone in population A is better off than everyone in population B, then A is better than B. Moreover, as Arrhenius has shown, there are theorems with logically weaker and intuitively even more compelling conditions.[17]

Impossibilities such as these are the challenges that motivate the Worry. One type of response to this challenge that we will set aside here is to offer a purported philosophical *resolution* to the challenge of the Repugnant Conclusion. Most of these purported resolutions argue that the Repugnant Conclusion should simply be accepted as true. For example, Hare (1988); Huemer (2008); J. L. Mackie (1985); Tännsjö (2002), and Gustafsson (forthcoming) have all offered arguments in favour of endorsing the Repugnant Conclusion, because of various arguments that the apparent repugnance of the conclusion is illusory or based on misunderstanding. One drawback with this resolution is that the theorems with logically weaker

[15] See Arrhenius (forthcoming), (2000a); Arrhenius, Ryberg, & Tännsjö (2014).
[16] For theorems with logically weaker and intuitively even more compelling conditions, see Arrhenius (forthcoming), (2000a), (2001), (2011).
[17] See, e.g., Arrhenius (forthcoming), (2000a), (2001), (2011).

conditions are not based on avoidance of the Repugnant Conclusion but on the intuitively more compelling **Very Repugnant Conclusion**: For any perfectly equal population with very high positive welfare, and for any number of lives with very negative welfare, there is a population consisting of the lives with negative welfare and lives with very low positive welfare which is better than the high welfare population, other things being equal.[18]

More recently, Budolfson & Spears (2018c) have offered an alternative type of resolution of the Repugnant Conclusion. They argue that Parfit's original example of the Repugnant Conclusion should be understood as describing only a proper subset of instances of the Repugnant Conclusion, and that the full set of instances of the Repugnant Conclusion should be understood to include a broader set, including cases in which there is a base population that is unaffected by the choice between a larger or a smaller population.[19] Given their more general characterization of the Repugnant Conclusion, they prove that all of the most commonly discussed aggregative welfarist population axiologies imply at least one instance of this unrestricted Repugnant Conclusion. They then argue that because the Repugnant Conclusion so understood is a problem for all of the most commonly discussed welfarist axiologies, it can no longer be reasonable to assume that a plausible axiology must avoid it.

We set aside these purported solutions in this paper. The problem we focus on is what the upshot of the population ethics literature is for policy on the assumption that there is no resolution to the challenges of population axiology at hand.

3. First Deflationary Response: Axiologies May Agree about Climate Change

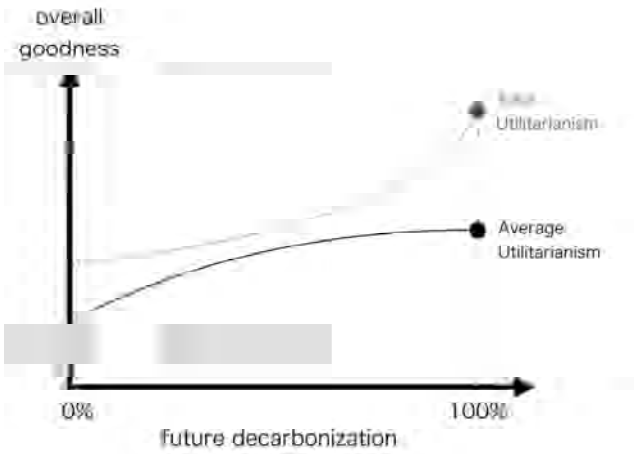
The open theoretical questions of population axiology only turn out to be a practical problem for a policy challenge if population axiologies sufficiently disagree about the best policy response to that challenge. To see how this could turn out not to be the case in connection with climate change, consider the toy illustrative example in

[18] See, e.g., Arrhenius (forthcoming), (2000b), (2011). For a detailed discussion of other problems with debunking arguments with regard to the Repugnant Conclusion, including Hare et al.'s arguments, see Arrhenius (forthcoming), ch. 3, (2000b).

[19] Budolfson and Spears' general characterization of the Repugnant Conclusion including instances with non-zero base populations is comparable to Arrhenius' Strong Quality Addition Principle (Arrhenius (forthcoming), (2000b)), which is violated by both Total and Average Utilitarianism (and some other population axiologies). Arrhenius draws, however, a different conclusion from this result, namely that the Strong Quality Addition Principle should be rejected as an adequacy condition since it rules out too many axiologies in one fell swoop and thus is in that sense too strong.

Figure 3. The figure plots a stylized version of the sort of climate policy decision considered by William Nordhaus' Integrated Assessment Models.

Figure 3. Two population axiologies recommend the same "corner solution" to optimal decarbonization



If figure 3 correctly described the full climate policy problem, then the Worry could be false, even though the candidate population axiologies differ. In the figure, the ethical question under consideration is what future decarbonization rate should be achieved: 100%, 0%, or some other optimum in between? The recommendations of two population axiologies are considered. These give different evaluations of different options. Total Utilitarianism rises convexly as the decarbonization rate increases; Average Utilitarianism rises only concavely. Thus, Average Utilitarianism thinks that a decarbonization rate of 90% would be only slightly worse than 100%, but Total Utilitarianism thinks 90% would be much worse than 100%.

Note that Average and Total Utilitarianism even have different *scales* for goodness: neither their lowest level of goodness nor their highest levels of goodness are the same number, and their evaluations cover ranges of different length. This is important because some responses to normative uncertainty — such as Expected Moral Value — recommend an average or expectation over alternative theories (Budolfson & Spears (2018a); Bykvist (2017); Bykvist, MacAskill, & Ord (2019); Greaves & Ord (2017); Hedden (2016)). This moral-expectation approach has found difficulty in the need to compare evaluation quantities across theories, but that problem is not relevant in the case of Figure 3, because the two axiologies agree on the optimum.

The point of Figure 3 is that both Average and Total Utilitarianism recommend the same *corner solution*. In optimization, a "corner solution" is when the optimal policy is equal to a boundary constraint. Because Average and Total Utilitarianism both recommend full decarbonization, in this example, there is no *practical* disagreement between them, only *theoretical* disagreement. Whether or not actual climate policy is well-described by figure 1 is substantially an empirical question (concerning economics, demography, climate science, etc.), although also a normative one (because different losses, such as of life and wealth, must be aggregated). However, it is not implausible that actual climate policy questions could be resolved by dominance — that is to say, by agreement across candidate axiologies. For example, if we are confident that a particular set of future lives would be full only of terrible suffering and thus not worth living, and if by preventing those lives from occurring we prevent some harmful carbon emissions, and if furthermore we know these are the only relevant considerations, then all plausible population axiologies recommend not creating those lives.

Although that example was fanciful, another might be quite realistic (see Scovronick et al. (2017) for detailed evaluation of the following). Consider investments in human development in developing countries, with a special focus on women's social status and the education and well-being of girls. This would have a range of likely consequences, which we can assume for hypothesis that we know with certainty (which would be confidence beyond the actual reach of social science):

- The women who receive the program and the lives lived by other people in their places and times would be better: an increase in the near-term average.
- Long-term average well-being would be improved by reduced climate change and by accelerated economic development.
- Some 21st century lives that would have been worth living would not be lived, because of empowered young women choosing to reduce their fertility. (Under Total Utilitarian-like theories, this would be a social cost.)
- Because of the reduced threat of climate change, the expected number of future good lives lived increases by more than the number of 21st century lives reduced.

In this case, the total expected number of lives lived would increase, average well-being would increase within every time period, and average across-time well-being

would increase because the average human would live later in historical time. Moreover, it is not implausible that the welfare of the worst-off lives would be higher (a property that matters to some egalitarian views), although this was not specified above. So, according to every plausible axiology in the literature and more — including Average utilitarianism and related views, Total Utilitarianism and related views, maximin, and others — implementing the human development policy is recommended, in expectation. The upshot is that we can know whether to implement the policy without knowing the correct population axiology, and also without a general solution to moral uncertainty. In this case, the Worry would be deflated.

More generally, other practical policy questions that are commonly taken to hinge on the choice of population axiology may be resolved by similar dominance arguments or corner solutions.[20] This would depend on social, economic, and scientific facts. For example, some have argued that an implication of Total Utilitarianism is that substantially more resources should be invested in preventing human extinction (Beckstead (2013); Bostrom (2013)). However, it may be that commonly-discussed policy options (such as asteroid deflection) offer a small marginal benefit of further investment as compared to merely pursuing standard economic growth, technological progress, and human development. The reason being that such standard policies would have large *co-benefits* against existential risk, perhaps because war of mass destruction or resistant, pandemic infectious disease would be less likely, or because survival-promoting technologies would be invented. If so, both Average and Total Utilitarianism would recommend serious investment in thoughtful, long-term human development, economic growth, and technical progress: Average Utilitarianism because it increases average well-being, and Total Utilitarianism because it does this while also offering the co-benefit of promoting survival. To be sure, this would not be the set of policies that humanity is currently pursuing, but it would not be a major reallocation into activities that only have the benefit of reducing existential risk, and nor would it turn on the choice of population axiology.

Of course, it may be that the climate policy menu under consideration does not yield one dominating option. Also, there could be additional considerations, such as bounded political capital. If political capital is scarce, a politician who needs to

[20] One exception to this possibility is the welfare of non-human animals. The number and well-being of nonhuman animals is generally governed by ecological forces such as natural selection, to a greater extent than the number and well-being of humans, which is regulated, in part, through complex technology and culture. In many cases, the implication of this fact may be that the average well-being of non-human animal species is kept within a narrow species-specific range, while adjustment to changing conditions occurs in population size (on the extensive rather than the intensive margin, in economists' language). If so, Average and Total Utilitarianism, as extended to non-human animals, may give very different recommendations. See Hsiung & Sunstein (2006), and Budolfson & Spears (2018b) for more on climate and non-human animals.

compromise across politically linked issues (such as climate policy and domestic health care or tax policy) may care about *how much worse* 95% would be than 100%, which cannot be settled by this sort of dominance-identification procedure. Still, this is a promising avenue for further research that should be pursued in light of the impossibility theorems in population axiology.

Recently, there have been attempts to resist this conclusion while holding on to PAC. Some (Roberts, Voorhoeve, and Fleurbaey) reject *Well-being entails being* and claim that non-existence does not preclude being better off (or worse-off).[21] Others (Adler, Arrhenius, Rabinowicz, Holtug, Johansson) instead reject *Better-for entails better-off* and claim that existence can be better for a person than non-existence even though the person would not be better off existing than not existing. A third option is to deny (i), i.e., deny that there are any non-identity cases, because one thinks that in all the worlds in which a person is not conceived, she still exists as a *merely possible* person, who has wellbeing. I shall argue that none of these ways of blocking the argument works. This leaves PAC itself as the only remaining culprit.

4. Second Deflationary Response: Bounded Population Principles

The Repugnant Conclusion --- and especially the search for a sensible population-sensitive social welfare function that does not imply the Repugnant Conclusion --- has been a central focus of the population ethics literature since Parfit (1984) introduced it. For example, Arrhenius, Ryberg, & Tannsjö (2014) has called it "one of the cardinal challenges of modern ethics" and Greaves (2017) introduces the Repugnant Conclusion as "the key objection" to Total Utilitarianism and related views. Because most of the literature on population axiology takes it as an adequacy condition that an acceptable social welfare function should not imply the Repugnant Conclusion, researchers have proven that many social welfare functions, in addition to total utilitarianism, imply the Repugnant Conclusion if the populations being evaluated can be unboundedly large. As noted above, Arrhenius (2000a), (2000b) presents an impossibility theorem that proves that no social welfare function can escape implying the Repugnant Conclusion, if the function is defined for unboundedly large population and has desirable --- and plausibly ethically necessary --- properties. Such properties are formalized as axioms for Arrhenius' theorems.

[21] Roberts (2015) does not defend PAC, but a weaker principle she calls 'the person-based intuition', according to which an outcome A is worse than an outcome B only if A is worse for someone than some alternative outcome Z, where Z need not be identical to B. However, she would have to defend PAC, if it is restricted to cases where A and B are the only available outcomes.

These are impressive and rigorous philosophical results. But what are the implications for policy analysis? Do these results show that the assumptions of many leading policy analyses are illegitimate, as suggested by the quotes above from IPSP and John Broome? More generally, how should policy analysis respond to these results? Arrhenius notes that one response could be a thoroughgoing scepticism or paralysis. However, he is much more enthusiastic about the possibility of a deflationary response: namely, to "try to find a way to explain away the relevance of the [Repugnant Conclusion and associated impossibility] theorem for moral justification."[22]

Our goal in this section is to articulate another deflationary response to the impossibility theorems to the effect that policy analysis can in some cases legitimately ignore them and the Repugnant Conclusion when that analysis applies to bounded problems, as Arrhenius's impossibility theorems assume unboundedness. We show that unlike unbounded cases, in bounded cases that are relevant to policy analysis, it is indeed possible to identify an axiology that captures all of the intoitions that support Total Utilitarianism while also avoiding the Repugnant Conclusion. This shows that it may be possible to endorse both the intuitions that motivate Total Utilitarianism and the intuition that tells against accepting the Repugnant Conclusion. The idea is that there might be a mere *appearance* of conflict between these intuitions that arises from taking our intuitions about the realistic range of cases relevant to policy as also extending to cases in the unbounded penumbra.

In other words, this second deflationary response to the Worry exploits the possibility of interpreting the intuitively compelling axioms of population ethics as restricted to a bounded domain.[23] An adequacy condition to avoid the Repugnant Conclusion on unbounded space has no implications for such a family of bounded axiologies. As we detail below, in our formal argument, our approach is not to reject that populations can be unboundedly large; instead, we propose bounded *axioms* that, in some cases, apply to only some of the space of possible populations.

4.1 Axiology with population size bounds

The practically relevant set of policy options that humanity will ever face is a bounded set, along many dimensions. This is partly because the set of practically

[22] Arrhenius (forthcoming), ch. 13, (2000b), ch. 12.

[23] Shiell (2008) offers a formal proof of an intuition (related to a point made by Parfit (1984), pg. 387), namely that within a truncated domain, Total Utilitarianism need not imply the Repugnant Conclusion within that domain. In this way, Shiell's proof depends essentially on truncating the choice set. In contrast, our proof below does not truncate the choice set. Our axiological principles cover the entire choice set, fully specify how to rank all outcomes within a policy-relevant range, but do not fully specify how rank all outcomes beyond that range. Moreover, the principles also satisfy certain bounded analogues of the central population ethics desiderata involved in the impossibility theorems in the area.

relevant population sizes is bounded. This is true even if the possible values of social welfare are unbounded, in part because policy choices could only have boundedly large effects on individual welfare. In making the empirical observation that the set of practically relevant population sizes is bounded, we have in mind a very large upper bound. The upper bound could be much larger than the largest set that an expert predicts could ever be relevant. It is sufficient for our purposes, for example, that the bound be 10^{80}, which is an estimate of the number of atoms in the universe, or 10^{58}, which is the estimate of Bostrom (2013) of the number of simulated human lives that a superintelligence could create with the available energy in the universe. The lower bound on the policy relevant set of population sizes is the number of humans who already have ever been born.

In this vein, even outside of population ethics, practical policy analyses are untroubled by imaginable, *unbounded* marginal utilities or counts of small harms; in this section, we formalize that observation by weakening some axioms of population ethics to a bounded domain. We can consider axioms that only apply to a very large but bounded subset of the potentially unbounded complete, imaginable social choice set, and choose a family of axiologies that (a) satisfies attractive axioms defined over the bounded set and (b) has no implications about the Repugnant Conclusion. A requirement to avoid the Repugnant Conclusion has no implications for this bounded family of axiologies.

The purpose of axiomatic representation theorems is to rule in and rule out sets of functional forms. In general, a representation theorem permits a *family* of function shapes that leaves certain features unspecified. For an example in the context of axiologies, critical level generalized utilitarianism is consistent with concave or affine transformations of utility and with positive or zero critical levels; each of these combinations would have different normative implications. Similarly, a family of population-sensitive axiologies could leave unspecified how populations are evaluated outside of the bounded set. Such a family of axiologies would ignore the Repugnant Conclusion --- while fully specifying the social evaluation on the bounded set.

The literature has identified the following very general characterization of the space of a number of important aggregative welfarist axiologies:

$$W = g(n)[h(n^{-1} \sum_i f(x_i)) - h(f(a))],$$[24]

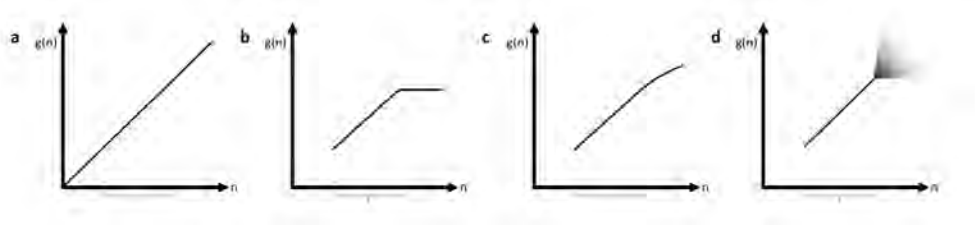
where:

[24] Compare Budolfson & Spears (2018c) and Greaves & Ord (2017).

- n is population size,
- x_i is the utility of person i,
- a is 0 or positive and is a critical level for adding a life to be a social improvement.
- The functions f, g, and h are all non-decreasing. If f and h are both the identity function, then we have utilitarianism. If f is concave and h is the identity function, then we have additively separable prioritarianism. If f is concave and $h = f^1$, we have a type of non-separable egalitarianism.

This general functional form is intended to clarify that the shape of g could be chosen independently of any combination of otherwise permissible features for the other elements of the function. It includes as special cases many axiologies in the literature, although not rank-dependent axiologies such as maximin or Zuber and Asheim's (2014) rank-dependent generalized utilitarianism, nor so-called person affecting theories.[25] In Total Utilitarianism g is linear; in Average Utilitarianism g is constant; and in Ng's Theory X' g is concave. Below, we will use the term "totalist" to refer to the family of theories according to which g is linear.

Figure 4. Families of social evaluations that cohere with totalist axioms on the bounded set



Note: Curly braces on the horizontal axis note the finite bounded set.

Figure 4 illustrates a possibility for g that is the focus of this section of the paper: a family of functional forms for g could be chosen that *fully specifies g* on the bounded policy-relevant set, while avoiding the Repugnant Conclusion and *taking no stand* on the shape of g outside the bounded set. Functional forms **a**, **b**, **c**, and **d** would rank policy options over the practically relevant set identically, for any given specifi-

[25] For a general discussion of the latter, see Arrhenius (forthcoming).

cation of *f*, *h*, and *a*. Form **a** matches Total Utilitarianism, if *f* and *h* are the identity function. Forms **b**, **c**, and **d** are blank at populations smaller that the bounded choice set, to illustrate that they do not make assumptions about how to rank populations this small. It is not essential to our argument that the bounded set have either a zero or a positive lower bound: the possibility of a lower bound greater than zero represents the minimum on policy-relevant population sizes due to the fact that billions of humans have already been born.

Forms **a**, **b**, and **c** have different implications for the Repugnant Conclusion, and may or may not invoke other undesirable properties outside of the practically relevant set. Form **d** is not a fully specified function form, but is merely a representation of the possibility of a decision-maker remaining uncertain about options outside of the bounded set. The existence of functional forms **a**, **b**, and **c** and of the options in **d** tells us that a climate policy-maker could say:

> Because over the practically relevant set of policy options I am both attracted to totalist intuitions (or axioms), and I am fully comfortable with a generalized total social welfare function; and because this practically relevant set is bounded, I should make policy according to any of **a**, **b**, **c**, or **d**. I remain troubled by the Repugnant Conclusion, but that can be a problem for future research, because it does not threaten my conviction about how policy options should be ordered in the practically relevant set of policy options.

Of course, someone with less totalist intuitions, for example someone who leans more toward Average Utilitarianism, wouldn't be able to say this. Likewise for theories that do not fall under the general characterisation above, such as rank-order theories and person affecting theories.[26] Still, it shows that restricting the applicability of the axioms to bounded sets opens up for convergence on policy recommendations for a number of different theories.

4.2 Possibility Proof for Escaping the Repugnant Conclusion while Satisfying Bounded versions of Population Ethics Desiderata

The graphical examples of the prior section suggest a route to avoiding the Repugnant Conclusion. In this section, we prove that this is possible by adopting a plausible set of axioms: namely, bounded versions of familiar axioms.

For example, in one of his pioneering informal results, Parfit (1984) makes use

[26] For a discussion of the latter family, see Arrhenius (forthcoming), (2000b).

of the controversial (since it makes it easy to derive the Repugnant Conclusion) Mere Addition Principle:

> **Mere Addition:** An addition of people with positive welfare does not make a population worse, other things being equal.[27]

This axiom could be weakened to:

> **Bounded Mere Addition:** An addition of people with positive welfare does not make a population worse, other things being equal, if each population (with and without the addition) is within the bounded domain.

One could similarly modify other adequacy condition axioms such as Arrhenius' Non-Sadism Condition to a Bounded Non-Sadism Condition, and the Egalitarian Dominance Condition to a Bounded Egalitarian Dominance Condition. In each case, the modified axiom would reflect an analogous axiological intuition as the original axiom, but with the restriction that it only applies to comparisons of populations within the bounded set. Such bounded axioms would simply make no claims about ranking populations outside of the bounded set. Relatedly, but outside of an axiomatic framework, one could assess the constructive argument that Broome (2004) presents for generalized, Critical-Level Total Utilitarianism, but --- unlike Broome --- only assess and apply the argument while considering populations within the bounded set.[28]

Would such bounded axioms be intuitively compelling? Because they are logically weaker than their unbounded counterparts, they must be at least as compelling. The impossibilities of population ethics are only interesting because the original axioms are compelling. Anyone who agrees with the original axioms will also agree with these, which are weaker: they make the same claims about fewer cases. And they may attract the new support of cautious evaluators who are hesitant to make axiomatic claims about unbounded populations.

In particular, consider a social evaluator who accepts the axiom of a complete and transitive social order for all populations, and accepts anonymity and same-number Pareto for all populations, but then accepts only the Bounded Mere

[27] See also Blackorby, Bossert, & Donaldson (2005), Arrhenius (forthcoming), (2000b). Like many contributors to the debate, Arrhenius and Blackorby et al. rejects the Mere Addition Principle as an adequacy condition for a satisfactory population axiology.

[28] Of course, a more substantive axiology such as Critical-Level Total Utilitarianism could still have unintuitive violations of other bounded conditions; for example, Critical-Level Total Utilitarianism violates a Bounded Non-Sadism that modifies the Non-Sadism axiom to only apply to the bounded set.

Addition and similarly modified and bounded versions of Separability and the other axioms that Blackorby & Donaldson (1984) demonstrate entail generalized Critical Level Total Utilitarianism. Such a set of axioms would entail a family of social welfare functions – each same-number utilitarian – where g is increasing and linear over the bounded set, and could have any shape outside of the bounded set (perhaps disciplined by further continuity axioms). In particular, the resulting axiologies need not be separable outside of the bounded set. Such bounded axioms would also rule out a positive critical level within the bounded set, due to Bounded Mere Addition. The modified axioms would provide a principled motivation for the social evaluator to use this family of social welfare functions. Such an axiology would be sufficient for a climate IAM and to answer any question posed by climate ethics, and the Repugnant Conclusion is not entailed.

More broadly, we now prove:

> **Possibility Theorem for Bounded Axiologies:** There exist complete welfarist axiologies that satisfy the Bounded Dominance, the Bounded Addition, and the Bounded Minimal Non-Extreme Priority Principles and avoid the Repugnant, the Bounded Sadistic, and the Bounded Anti-Egalitarian Conclusion.

The proof is by example. Forms **b** and **c** from Figure 4 satisfy the theorem, as does any form of W in which h and f are the identity functions, g is the identity function on the bounded set (as in Total Utilitarianism), and g is everywhere non-decreasing and is bounded above outside the bounded set. At very large population sizes outside of the bounded set, this family of axiologies would imply the (unbounded) Sadistic Conclusion, just as Ng's Theory X' does – but that is no contradiction, because the Possibility Theorem only requires avoiding the Sadistic Conclusion in the bounded set. Note that bounded Average Utilitarianism (g is constant in the bounded set) is not an example consistent with the Possibility Theorem because it does not satisfy avoiding even the Bounded Sadistic Conclusion; nor does Theory X', if g is concave within the bounded set.

A worry, however, is that the impossibility theorems might reappear over a bounded domain by further reformulating the adequacy conditions to take into account that we are now dealing with a bounded domain. Such reformulations can be done in multiple ways, one straightforward example is as follows:

> **Bounded Repugnant Conclusion I:** In the bounded domain, for any population consisting of people with very high positive welfare, there is a better population in which everyone has a very low positive welfare, other things being equal.

Rather trivially, this cannot be an implication of axiologies that verify the Possibility Theorem above. Consider, for example, the largest population size within the bounded domain, and assume each member of that population has a very high welfare. Because this involves the largest population size within the domain, there cannot be a population with much lower welfare that is better.

However, there are other reformulations of the Repugnant Conclusion that are not as easily avoided in the bounded domain. Here is one example:

> **Bounded Repugnant Conclusion II:** In the bounded domain, there are very large populations consisting of people each with very high positive welfare for which there are better populations in which everyone has a very low positive welfare, other things being equal.

The idea behind the Bounded Repugnant Conclusion II is the intuition that if a population is sufficiently big and everyone enjoys very high welfare, then such a population is better than each of the populations with only very low positive welfare in the domain. This intuition is one candidate for being the main intuition behind the counterintuitiveness of the original Repugnant Conclusion (recall that Parfit formulated it in terms of "any possible population *of at least ten billion people*"[29]).

Along this line, it could be further argued that what is fundamental to repugnance is the existence of a **Large Quantity-Quality Tradeoff** – meaning, a case where a large increase in quantity is allowed to compensate for a large decrease in quantity, or the reverse. According to this take on the Repugnant Conclusion, unboundedness is not essential to repugnance. This raises the important question of what is essential to the repugnance of the Repugnant Conclusion, and how many versions or instances there may be. As it is sometimes expressed, there can be various instances of the Repugnant Conclusion (Parfit (2016)). If so, perhaps a satisfactory population axiology should not imply any instances of it.

Depending on the size of the domain, the size of the very large populations, and on what the difference is between lives with very high and very low welfare, Bounded Total Utilitarianism might imply Bounded Repugnant Conclusion II. For example, let's assume that a life with very high welfare is at least 100 times better than a life with very low positive welfare and let's use Bostrom's estimate, mentioned above, of 10^{58} simulated human lives as an upper bound on the size of possible populations. It follows from Bounded Total Utilitarianism that there is a very high welfare level such that for any population up to size 10^{56} enjoying this level, there is a better very low welfare population in the domain. So, according to Bounded Total Utilitaria-

[29] Parfit (1984), p. 388, emphasis added.

nism, a population with lives barely worth living would be better than an enormous population with very high individual quality of life. And given that an intuitively sufficiently large population with very high welfare is smaller than 10^{56}, which seems intuitively compelling (compare Parfit's specification of "at least 10 billion people"), Bounded Total Utilitarianism implies the Bounded Repugnant Conclusion II in this domain.

One can, of course, argue for other smaller upper bounds on the size of possible populations and for other differences between very high and very low positive welfare lives. However, what this shows is that the unbounded scope of the classical Repugnant Conclusion is not needed to produce extreme quantity-quality trade-offs. More importantly, it shows that there may be impossibility theorems looming even in the bounded domain with the adequacy conditions from the unrestricted domain appropriately adjusted. Of course, this has to be appropriately shown by proving such theorems.

The mere fact that some set of axioms is impossible to combine is not sufficient, of course, for an important challenge to climate policy-making. The involved conditions also have to be intuitively compelling. As the example above hints at, these conditions might or might not be sufficiently compelling depending on what one takes to be the main intuition behind classical unbounded conditions. Hence, the results we get when restricting population ethics to a bounded domain raises new and important questions that need to be further investigated: Is the implication of Bounded Repugnant Conclusion II sufficiently counterintuitive to work as an adequacy condition for a satisfactory population ethics? Might it even capture the main intuition behind the counterintuitiveness of the original Repugnant Conclusion? Or is unboundedness an essential part of the counterintuitiveness of the Repugnant Conclusion?

More broadly, this result suggests asking why exactly the Repugnant Conclusion is counterintuitive. Is the quantity-quality trade-off involved in the Bounded Repugnant Conclusion II sufficiently similar to a general quality-quantity trade-off problem for every aggregative axiology (see Budolfson & Spears (2018c), discussed above) to make it unsuitable as a condition on theory choice with respect to aggregative axiologies?

Ultimately, we need to scrutinize more carefully the source of the counterintuitiveness of the original Repugnant Conclusion to know whether it will carry over to the bounded domain. Moreover, could the force of bounded impossibility theorems be weakened by finding good reasons to restrict the upper bound on the domain further? And will the further assumptions that seem to be needed for bounded theorems, such as assumptions regarding the possible size of the involved populations, the difference between very high and very low positive welfare, and the

measurement of welfare (in the above example we assumed a ratio scale which isn't necessary for the unbounded theorems) open up for ways of escaping the theorems that are not available in the unbounded domain? This is an important but neglected area of research in population ethics which the second deflationary response puts focus on.

5. Conclusion

Policy analysis requires an axiology, population dynamics are important to climate change, and there is radical disagreement among experts about population axiology (Arrhenius (forthcoming), (2000a), (2000b), (2001), (2011)). Does this state of affairs limit our ability to know how to respond to climate change? Although several prominent voices have voiced this Worry, we suggested that it is not obviously well-founded, and we have highlighted two possible deflationary responses. In the first, we noted that many important policy questions are likely to be subject to simple, cross-theoretical dominance resolutions, as illustrated by a corner solution to an optimization problem. In the second deflationary response, we observed that the intuitions that support the axioms that lead to the Repugnant Conclusion also support the axioms in the bounded case while avoiding the Repugnant Conclusion. Because any real-world policy question is a question about a bounded population domain (even if potentially very large in quantity), we can adopt these axioms for purposes of policy in their modified bounded form.

We also noted some important limitations and possible problems for these deflationary strategies. Regarding the first deflationary response, we noted that the climate policy menu under consideration may not yield one dominating option. Moreover, there could be additional considerations, such as bounded political capital, which could complicate the issue such that it cannot be settled by the suggested dominance-identification procedure, or could simply the issue by further reducing the practical space of policy options to those in which many axiologies agree.

Regarding the second deflationary response, there is the worry that the impossibility theorems might reappear over a bounded domain when the classical adequacy conditions are appropriately adjusted for the bounded domain. An important challenge highlighted by considering the Repugnant Conclusion on a bounded domain is the need to identify exactly what constitutes the main counter-intuitiveness of the Repugnant Conclusion and whether it carries over from the unbounded to the bounded domain (or, perhaps, to any other domains). This is a neglected but important area for further research in light of the impossibility

theorems in population axiology on unbounded domains and the possibility theorem above on bounded domains.

In the meantime, we need not overstate the *practical* importance of the Repugnant Conclusion and other challenging problems in population ethics as we seek to cope with important challenges for the future of humanity. As we have shown, scepticism and paralysis are not yet warranted, as there are promising deflationary responses to the impossibility theorems and strategies for gaining consensus given disagreement for practical policymaking. Policy analysis may not need to wait for greater consensus in population ethics.[30]

References

Anglin, W. (1977). The Repugnant Conclusion. *Canadian Journal of Philosophy, 7*(4), 745–754.

Arrhenius, G. (forthcoming). *Population Ethics: The Challenge of Future Generations*. Oxford University Press.

Arrhenius, G. (2000a). An Impossibility Theorem for Welfarist Axiologies. *Economics and Philosophy, 16*(02), 247–266.

Arrhenius, G. (2000b). *Future Generations: A Challenge for Moral Theory*. Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2:170236

Arrhenius, G. (2001). What Österberg's Population Theory Has in Common With Plato's. In *Omnium-gatherum. Philosophical Essays Dedicated to Jan Österberg on the Occastion of his Sixtieth Birthday* (Vol. 50, pp. 29–44). Uppsala: Department of Philosophy, Uppsala University: Uppsala Philosophical Studies.

Arrhenius, G. (2011). The Impossibility of a Satisfactory Population Ethics. In H. Colonius & E. N. Dzhafarov (Eds.), *Descriptive and Normative Approaches to Human Behavior, Advanced Series on Mathematical Psychology* (pp. 1–26). World Scientific Publishing Company.

Arrhenius, G., Ryberg, J., & Tännsjö, T. (2014). The Repugnant Conclusion. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2014). Retrieved from http://plato.stanford.edu/archives/spr2014/entries/repugnant-conclusion/.

[30] Thanks to Andrea Asker, Drew Burd, Krister Bykvist, Tim Campbell, Diane Coffey, Iwao Hirose, Gerald Lang, Melissa LoPalo, Kevin Kuruc, Tristram McPherson, Josh Petersen, Shlomi Segall, Sangita Vyas, and the audiences at Paris School of Economics, the Australian National University, and the Institute for Futures Studies.

Asheim, G. B., & Zuber, S. (2014). Escaping the repugnant conclusion: Rank discounted utilitarianism with variable population. *Theoretical Economics*, *9*(3), 629–650. https://doi.org/10.3982/TE1338.

Beckstead, N. (2013). *On the Overwhelming Importance of Shaping the Far Future*. New Brunswick, NJ.

Blackorby, C., Bossert, W., & Donaldson, D. (1995). Intertemporal Population Ethics: Critical-Level Utilitarian Principles. *Econometrica*, *63*(6), 1303–1320. https://doi.org/10.2307/2171771.

Blackorby, C., Bossert, W., & Donaldson, D. J. (2005). *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*. New York: Cambridge University Press.

Blackorby, C., & Donaldson, D. (1984). Social Criteria for Evaluating Population Change. *Journal of Public Economics*, *25*(1–2), 13–33. https://doi.org/10.1016/0047-2727(84)90042-2.

Bostrom, N. (2013). Existential Risk Prevention as Global Priority. *Global Policy*, *4*(1), 15–31. https://doi.org/10.1111/1758-5899.12002.

Broad, C. D. (1979). *Five Types of Ethical Theory* (1 edition). London: Routledge.

Broome, J. (1992). *Counting the Cost of Global Warming*. Cambridge: The White Horse Press.

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

Broome, J. (2010). The most important thing about climate change. *Why Ethics Matters*, 101.

Broome, J. (2012a). *Climate Matters*. Norton.

Broome, J. (2012b). *Climate Matters: Ethics in a Warming World*. Retrieved from https://books.google.com/books?hl=en&lr=&id=RjrYYEk8GYQC&oi=fnd&pg=PA1&dq=%22to+their+clearest+and+most+compelling+essences.+Our+hope+is%22+%22the+Cost+of+Global%22+%228:+The+Future+versus+the%22+%22series+will+broaden+the+set+of+issues+taken+up+by+the+human%22+&ots=ctUEq1iVaP&sig=F6Y0B7587dZBLxnbtvRb9DPZUXc.

Budolfson, M., & Spears, D. (2018a). *An impossibility result for decision-making under normative uncertainty*. mimeo.

Budolfson, M., & Spears, D. (2018b). *Methods for quantifying animal wellbeing and estimating optimal tradeoffs against human wellbeing — and lessons for axiology, including new arguments for separability*. mimeo.

Budolfson, M., & Spears, D. (2018c). *Why the Repugnant Conclusion is Inescapable*. mimeo.

Bykvist, K. (2017). Moral Uncertainty. *Philosophy Compass*, *12*(3), 1–8.

Bykvist, K., MacAskill, W., & Ord, T. (2019). *Moral uncertainty*. Oxford: Oxford University Press.

Greaves, H. (2017). Population Axiology. *Philosophy Compass*, *12*(11), 1–15.

Greaves, H., & Ord, T. (2017). Moral Uncertainty About Population Axiology. *Journal of Ethics and Social Philosophy*, *12*(2).

Gustafsson, J. E. (forthcoming). Our Intuitive Grasp of the Repugnant Conclusion. In Arrhenius, Gustaf, K. Bykvist, T. Campbell, & E. Finneron-Burns (Eds.), *The Oxford Handbook of Population Ethics*. Oxford: Oxford University Press.

Hare, R. M. (1988). Possible People. *Bioethics*, *2*(4), 279–293.

Hedden, B. (2016). Does MITE make right? In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vols. 1–11, pp. 102–135). Oxford University Press.

Hsiung, W., & Sunstein, C. R. (2006). Climate Change and Animals. *University of Pennsylvania Law Review*, *155*(6), 1695–1740.

Huemer, M. (2008). In Defence of Repugnance. *Mind*, *117*(468), 899–933. https://doi.org/10.1093/mind/fzn079

Mackie, J. L. (1985). Parfit's Population Paradox. In J. Mackie & P. Mackie (Eds.), *Persons and Values* (pp. 242–248). Oxford: Oxford University Press.

McMahan, J. (1981). Review: Problems of Population Theory. *Ethics*, *92*(1), 96–127.

McTaggart, J. M. E. (1927). *The Nature of Existence*. Cambridge.

Narveson, J. (1967). Utilitarianism and New Generations. *Mind, 76*(301), 62–72.

Ng, Y.-K. (1989). What Should We Do About Future Generations? *Economics and Philosophy*, *5*(02), 235–253. https://doi.org/10.1017/S0266267100002406

Pachauri, R. K., Mayer, L., & Intergovernmental Panel on Climate Change (Eds.). (2015). *Climate change 2014: synthesis report*. Geneva, Switzerland: Intergovernmental Panel on Climate Change.

Parfit, D. (1982). Future Generations: Further Problems. *Philosophy & Public Affairs*, *11*(02), 113–172.

Parfit, D. (1984). *Reasons and Persons* (1991st ed.). Oxford: Clarendon.

Parfit, D. (1986). Overpopulation and the Quality of Life. In P. Singer (Ed.), *Applied Ethics* (1 edition, pp. 145–164). Oxford: New York: Oxford University Press.

Parfit, D. (2016). Can We Avoid the Repugnant Conclusion? *Theoria*, *82*, 110–127.

Scovronick, N., Budolfson, M. B., Dennig, F., Fleurbaey, M., Siebert, A., Socolow, R. H., … Wagner, F. (2017). Impact of population growth and population ethics on climate change mitigation policy. *PNAS*, *114*(46), 12338–12343.

Shiell, L. (2008). The Repugnant Conclusion and Utilitarianism under Domain Restriction. *Journal of Public Economic Theory*, *10*(6), 1011–1031.

Sider, T. R. (1991). Might Theory X Be a Theory of Diminishing Marginal Value? *Analysis*, *51*(4), 265–271.

Sidgwick, H. (1907). *The Methods of Ethics*. London: Macmillan.

Spears, D. (2019). The Asymmetry of population ethics: experimental social choice and dual-process moral reasoning. *Economics and Philosophy*.

Tännsjö, T. (2002). Why We Ought to Accept the Repugnant Conclusion. *Utilitas*, *14*(03), 339–359.

Appendix: A Smoothness Axiom and a New Argument for Total Utilitarianism Full scales

One response to the argument in Section 4 of the paper would be to agree that the modified axioms in their bounded versions capture *some* of our important intuitions, but not *all* of them, because there is a specific intuition that is omitted: that axiology is infinitely continuous. Consider the case in which a family of axiology is chosen, based on axioms some bounded and some unbounded, such that a social welfare function of form W is chosen, with the additional properties that:

- Bounded separability is assumed in social evaluation, so that the social welfare function can be written as a function of two variables: $\widehat{W} = g(\bar{n})h(\tilde{x})$, where $\bar{n}$ is the expected size of the population and $\tilde{x}$ is the expectation of $f(x)$. Then, g and the other functions are functions of all real numbers (not just counting numbers).
- f and h are both identity functions, as in total or average utilitarianism or Theory X', so the expression simplifies to: $\widehat{W} = g(\bar{n})\,\bar{x}$, where $\bar{x}$ is average utility.
- g is the identity function on the bounded set, as in total utilitarianism, and is any non-decreasing function outside of the bounded set, so the Repugnant Conclusion is not logically entailed (and therefore may or may not be avoided).

This is the sort of family of social welfare functions that section 4 highlights as possible, but extended for illustration to the case of expectations, in order to cover real numbers (and not only counting numbers of people); this will not appeal to advocates of non-expected social evaluations.

Now consider the intuition that axiology should be infinitely continuous – an intuition that may appear as an experience of unease about the boundedness of axioms. We can formalize this axiom as:

> **Smoothness**: g is C^{∞}, which is mathematical notation for the property of a function in which each derivative is continuous everywhere.

For real-valued functions, the Smoothness axiom would imply that they are polynomials. Therefore, g must be the identity function everywhere, because it is the identity function in the bounded set. The upshot is that the bounded

assumptions above *plus the Smoothness axiom* imply that $\widehat{W}$ is expected Total Utilitarianism.[31]

The Smoothness axiom – and the intuitive response to the boundedness proposal that it captures – is therefore a new, constructive argument for Total Utilitarianism. With the smoothness axiom, $\widehat{W}$ implies the Repugnant Conclusion. Therefore, the Smoothness axiom introduces a new theoretical cost of avoiding the Repugnant Conclusion, in the context of the bounded axioms of $\widehat{W}$. If you find boundedness distasteful because you find infinite continuity to be a plausibly compelling property of axiology, then that intuition – in combination with other axioms – is a new argument counting in favour of Total Utilitarianism and acceptance of the Repugnant Conclusion. Of course, it can also be taken as a new impossibility theorem for those who accept smoothness, the bounded assumptions above, but not the Repugnant Conclusion.

[31] Thanks to Kevin Kuruc for suggesting consideration of this argument.

Mark Budolfson[1] & Dean Spears[2]

Population ethics and the prospects for fertility policy as climate mitigation policy

What are the prospects for using population policy as tool to reduce carbon emissions? In this paper, we review evidence from population science, in order to inform debates in population ethics that, so far, have largely taken place within the academic philosophy literature. In particular, we ask whether fertility policy is likely to have a large effect on carbon emissions, and therefore on temperature change. Our answer is no. Prospects for a policy of fertility-reduction-as-climate-mitigation are limited by population momentum, a demographic factor that limits possible variation in the size of the population, even if fertility rates change very quickly. In particular, a hypothetical policy that instantaneously changed fertility and mortality rates to replacement levels would nevertheless result in a population of over 9 billion people in 2060. We use a leading climate-economy model to project the consequence of such a hypothetical policy for climate change. As a standalone mitigation policy, such a hypothetical change in the size of the future population – much too large to be

[1] University of Vermont Gund Institute for Environment and Department of Philosophy and Australian National University.
[2] Corresponding author. Department of Economics and Population Research Center, University of Texas at Austin. Indian Statistical Institute, Delhi Centre. IZA Institute of Labor Economics. Institute for Futures Studies. dspears@utexas.edu. Dean Spears' research was supported by grant K01HD098313 and by P2CHD042849, Population Research Center, awarded to the Population Research Center at The University of Texas at Austin by the Eunice Kennedy Shriver National Institute of Child Health and Human Development. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

implementable by any foreseeable government program – would reduce peak temperature change only to 6.4°C, relative to 7.1°C under the most likely population path. Therefore, fertility reduction is unlikely to be an adequate core approach to climate mitigation.

*

1. Introduction

What does the threat of climate change mean for population policy? Much of the debate on this question in the literature has been focused on population ethics, and therefore has mainly involved dialogue within philosophy and related disciplines. *Population ethics* is a subfield of philosophy that asks when and whether an increase or decrease in the size of the population is a social improvement, or would be a good goal for policy to pursue. For example, one classic question in population ethics is whether policy-makers should try to maximize average well-being or total well-being. Some philosophers have worried that, because ethicists have not settled the theoretical questions of population ethics, and because population must be an important component of climate policy, policy-makers cannot yet know what should be done about climate change.

The purpose of this special issue is to broaden the dialogue between population ethicists and empirical demographers. In this paper, we present some demographic facts and arguments, with the goal of informing debates in population ethics about climate policy.[3] For example, a recent philosophy paper begins "In recent years increasing numbers of moral and political philosophers have argued than an adequate response to global environmental challenges, such as climate change, requires adopting policies that will either slow down global population growth or even reduce global population size."[4] But, empirically, is it correct that such policies would be feasible and would meaningfully influence climate change? Whether or not debates in population ethics are relevant to climate policy depends, in part, on empirical questions such as this one.

In fact, global population growth *already* has slowed down substantially from its historically exceptional peak rate of over 2% per year in the mid-20th century to around 1% now. The growth rate is expected to continue falling and converge to zero.

[3] The arguments in the paper build upon, and in some places summarize for an audience of population axiologists, the work of prior empirical demographers and social scientists, including Pritchett (1994), Connelly (2009), Lam (2011), O'Neill, et al. (2012), and Bradshaw and Brook (2014).
[4] This quotation is the opening of Caney (2019); we highlight this thoughtful paper here not to criticize it, but to illustrate the literature in which it participates.

Although the size of the population doubled in less than 40 years in the second half of the 20th century, it is projected to never double again (Lam, 2011). The size of the population is projected to peak shortly after the beginning of the 22nd century (Gerland, et al., 2014). After that, the size of the human population is projected to decline.

In what the UN calls "more developed regions" there are about 1.27 billion people now, and there are projected to be about 1.28 billion in 2100.[5] Asia, overall, is projected to move from 4.6 billion now only to 4.8 billion in 2100: in between, the population of Asia will peak and begin to decline. The outlier is sub-Saharan Africa, where fertility remains high, so population size is projected to quadruple from a little below 1 billion now to 4 billion in 2100.

Mid-20th century population growth was caused by a temporary excess of fertility over mortality: mortality fell quickly as a result of economic and human development and especially advances in public health, infectious disease, and sanitation. Fertility rates have taken a few decades to follow downwards. However, because mortality rates can only fall to zero and not below, this one-time transition cannot be repeated. Under any foreseeable demographic trajectory, the rapid population growth of the 20th century was a one-time event.

Therefore, the practical policy question is not whether population growth rates should decline (they already are) or whether there should be a peak size of the human population (there will be, in about 100 years or so). The practical policy question is whether steps should be taken to make the ongoing fall in fertility rates proceed *even more quickly*. If so, the region where fertility rates are high enough that there is scope, in principle, for faster decline is sub-Saharan Africa – where emissions per capita are currently low. Fertility is substantially shaped by the behavior and choices of women and families. Across developing countries and years, achieved fertility is highly correlated with intended fertility (Pritchett 1994). Although many historical fertility policies have been coercive and harmful (Connelly, 2009), we assume that readers of this paper are interested in voluntary or incentive-based policies that come at only low to moderate social cost to present generations. So, a plausible policy that substantially accelerated the decline in fertility rates would have to be implemented by states or institutions (especially in sub-Saharan Africa), to influence average concepts of *ideal or desired* fertility in those populations, rapidly enough to make a difference on the near-term timeline that is relevant to climate policy.

[5] Each projection in this section is from the median of the 2017 revision of the UN World Population Prospects.

We argue, based on empirical demography, that this is quantitatively implausible as a central tool for climate mitigation policy.[6] Fertility reduction, as a consequence of human development policy, may play a part in the response to climate change, but it is not the case that population size must, should, or can be the core of a sufficient climate mitigation policy. Moreover, if fertility policy comes at the political opportunity cost of pursuing other climate mitigation policies (here, we do not argue that it necessarily would), then a focus on population as the core of an emissions-reduction strategy is unlikely to succeed.

Section 2 situates our argument within the literature on population and climate policy. Section 3 introduces population momentum and the narrow scope for change in population growth rates in coming decades. We examine the consequences for climate outcomes of one particular hypothetical population policy that instantaneously changed fertility and mortality rates to replacement levels. As a separate observation, Section 4 observes that achieved fertility is highly correlated, across populations, with intended fertility. If fertility is high in some populations due to high desired fertility (and not, for example, due to unmet need for contraception) then fertility-reduction may be hard to achieve as a policy goal. Section 5 concludes.

2. Our argument, and its place in the literature

The purpose of this paper is to highlight some empirical facts about population science and climate change, and to connect those facts to debates in the population ethics literature. In this section, we begin merely by noting that there are many mechanisms listed in the literature by which population and climate policies may interact. What population ethics will have to say about a policy option depends, in part, on which mechanism is in question.

2.1 Our question, among mechanisms in the literature

Mechanism i. Effects of population size on emissions: Population-reduction policy as climate mitigation policy.

If there are more people, and if emissions per capita remain unchanged or similar, then the flow of emissions per time period would increase (O'Neill, et al., 2012). As a result, temperature change will be greater, all else equal. This fact has motivated a

[6] If readily available, we would endorse an adequate set of human development policies that have fertility reduction as a side-effect, or even some human development policies with a fertility-reduction goal as among a large group of "climate policy wedges" (Pacala and Socolow, 2004), each making their small contribution.

debate over whether policy should attempt to reduce the size of the near-future population, as a climate mitigation policy tool, to reduce carbon emissions.

Mechanism ii. Consequences of the exogenous size of the future population for optimal forward-looking mitigation policy.

Mechanism *ii* is a consequence of the size of the future population, in the special case where climate policy is being chosen to promote overall social well-being.[7] If there are more future people because of an exogenous difference in the population path, then more future people will be exposed to temperature changes, and therefore harmed by climate change. As a result, the harm done by carbon emissions today is greater, so the optimal level of present-day carbon emissions is lower. Scovronick, et al., 2017 and Budolfson, et al. 2018 show, using a leading climate-economy model, that this mechanism has a quantitatively important effect on optimal near-term emissions policy. Such an effect depends on the choice of social welfare function (that is to say, different theories in population ethics), because different social welfare functions incorporate the size of the future population in different ways.[8] To emphasize, under this mechanism the population path is not the core emissions-reduction strategy: Mechanism *ii* holds that there additionally exists a core emissions-reduction strategy, such as a carbon tax, and that the exogenous path of the future population changes how large that carbon tax ideally should be.

Mechanism iii. Effects of climate change on the size of the future population: Empirical.

Climate change – directly through temperature or indirectly through disease, drought, and other mechanisms – could change mortality rates, including for babies, and could change fertility rates.[9] Over the long run, climate change could cause large changes in the size of the human population, relative to a future in which temperatures stayed at pre-industrial.

---

[7] As Scovronick, et al. 2017 show, this mechanism operates for the selection of optimal policy and also for the selection of the optimal way of achieving a (potentially non-optimal) temperature target.

[8] If population size is exogenously held fixed, some theories in population ethics become identical to one another. For example, total utilitarianism, average utilitarianism that averages over all time, and Ng's (1989) Theory *X*' that incorporates all time would all three rank identically any set of policies in which population size is held constant. However, other social welfare functions in the literature, such as Asheim and Zuber's (2014) Rank-Dependent Generalized Utilitarianism, could disagree with these three, even in cases where population size is held constant.

[9] Some recent papers that document or project effects of temperature on mortality include: Sherwood and Huber (2010), Barreca (2012), Barreca, et al (2016), and Geruso and Spears (2018). See Spears (2019) for a review. In particular, Sherwood and Huber note that humans may not be able to survive exposure to combinations of heat and humidity that could plausibly occur under climate change, in which the human body would not be able to cool itself by sweating. Climate change might also increase the variance of the size of the future population (Spears, 2015).

Mechanism iv. Effects of climate change on the size of the future population: Population ethics.
Mechanism *iv* assumes that Mechanism *iii* is correct and that climate change will importantly change the size of the future population, such as causing fewer lives to be lived. Philosophers who study population ethics (or, in particular, population axiology, which focuses on how population size impacts rankings of goodness) ask whether the absence of such lives worth living counts as a social cost that policy should try to avoid. In other words, if it is true that climate change will cause some lives that would have been good to instead never be lived at all (because the people are not born), does this *consequence* of climate change count as a *cost* of climate change?

We have previously studied Mechanism *ii* in Scovronick, et al. (2017). We have investigated *iii* and *iv* in depth in Arrhenius, et al. (2019), and do not focus on them here. Mechanism *i* is the focus of this paper. Our objective is to bring facts and insights from empirical demography into the debates of population ethicists and others. In short: would fertility policy "work" as climate mitigation policy? Can policy-makers hope to limit temperature change through a strategy of reducing the size of the population.

2.1 Our question, among mechanisms in the literature

We consider Mechanism *i*: the policy consequences of the possible effect of the size of the population on emissions. In fact, throughout this paper we assume for the sake of argument that there is an effect on the size of the population in a near-term time period, under near-term technology, on the rate of carbon emissions. Nothing in this paper argues against the possibility of a large effect of the size of the future population on emissions. But what would this assumption, if true, imply for population policy?

Our question is not whether current fertility levels are optimal or whether a lower fertility level would be optimal (we take no stance here). Our question is also not whether it would be a desirable policy outcome for all women to have access to reproductive health care and broad social equality (it would be). Our question is: *given actual constraints on demographic change, governance, and policy-making attention, is fertility-reduction likely to be a quantitatively successful climate mitigation policy?* We argue that it is not.

Note that, because we are asking about states and related institutions, we do not take a position on the question whether parents have a right to have more than one child (Conly 2016) or whether it is wrong to create or not to create a particular life (Roberts 2019), or on any other question about one woman or family's procreative

behavior. To be clear: expanding access to reproductive health care and promoting human development objectives such as education and social equality are desirable goals. Moreover, these goals perhaps should receive even more resources than they otherwise would, in a tradeoff against some competing policy priorities (such as, for example, balancing government budgets), because of their implications for climate change.[10] However, reducing fertility should not be a first-order priority *as a substitute for other climate mitigation policies.* We also emphasize that our arguments depend on empirical premises: if there were a surprising opportunity to pursue a human development policy that reduced fertility while also making present-day people better-off and without crowding out more effective climate mitigation policy, then we would have no objection – but we expect that the effect of such a policy on temperature change would be small.

Our argument is based on three empirical observations, for which the body of this paper reviews empirical evidence:

> **Observation 1.** Population projections are highly certain over the coming decades, which is the time period when effective climate mitigation must occur.
>
> **Observation 2.** The regions where population projections are more uncertain, and therefore where fertility could fall, are places with low emissions per capita. In particular, uncertainty is highest in sub-Saharan Africa.
>
> **Observation 3.** On average, achieved fertility is highly correlated with intended fertility. So, policy to reduce fertility would have to change *intended* fertility. It is not fully understood why intended fertility remains high in sub-Saharan Africa.

So, our argument is a pragmatic one: intended as an emissions-reduction strategy, fertility policy by near-term governments is unlikely to have climate mitigation benefits that exceed the social costs, which include any opportunity costs of pursuing other climate mitigation strategies. We address two questions: what possible changes in the near-term size of the human population are plausible, and could actual governments plausibly achieve them at acceptable social costs? One seeming paradox might be an apparent contradiction between Observation 1 and Observation 3: how can the future path of the population be accurately projected, if fertility intentions remain poorly understood? The next section presents the explanation: population momentum.

[10] This is what we show in Scovronick, et al. (2017). Investments in human development that result in a very low population trajectory would substantially reduce the costs of climate mitigation policy, in a comparison of optimal climate mitigation policy under a very low population trajectory versus optimal climate mitigation policy under a business-as-usual population trajectory.

3. The tension between PAC and our intuitive judgements about non-identity cases

Population momentum is demographers' term for the fact that the size of the population would continue to increase even if the total fertility rate [11] (TFR) hypothetically instantaneously dropped from higher levels to replacement levels.[12] Population momentum occurs because today's baby girls will grow up to be women of reproductive age. So, if there are more girls at each pre-childbearing age than there currently are women at each childbearing age, then more babies than today will be born when those girls begin having children, even if fertility per mother is held constant. As a result, demographers can be highly certain that population in sub-Saharan Africa will continue to grow, even if the rate of decline over time in fertility intentions is suddenly accelerated. In particular, even if total fertility rates instantaneously changed to replacement levels in every country around the world (which would be an extreme outcome that no known policy could effect), the population in 2060 would still have about 9 billion people, or almost one-fourth more people than are alive today.

3.1 Global population uncertainty, relative to the urgency of climate mitigation

Figure 1 puts population possibilities in context by plotting possible population futures against the size of needed climate policy. Future population paths are taken from the 2017 UN World Population Prospects. For each year, for each projection, the graph plots the percent reduction in the size of the population, under that path, relative to population in that year under the UN's median path. Four population reduction paths are plotted. Two are for the UN's low 80% trajectory and two are for the UN's low 95% trajectory. These low trajectories are the bottom of 80% and 95% confidence intervals, respectively, for the future population. In other words, the UN projects that there is an 80% chance that the future population path will be within the low path and the high path (the high path is not used to produce this figure).

For each of these "low" paths, the figure plots the percent by which the low path is below the median path. In this way, the figure is intended to roughly quantify

[11] The *total fertility rate* of a cohort is the average number of children had over the course of a childbearing career by women who survive their entire childbearing career. Period TFRs are computed by assuming that a synthetic cohort experiences the age-specific fertility rates that prevail in a population in a given period.

[12] This portion of our argument builds upon related prior arguments by Bradshaw and Brook (2014), although they do not apply their observation to the climate-economy modeling that is the novel contribution of this section.

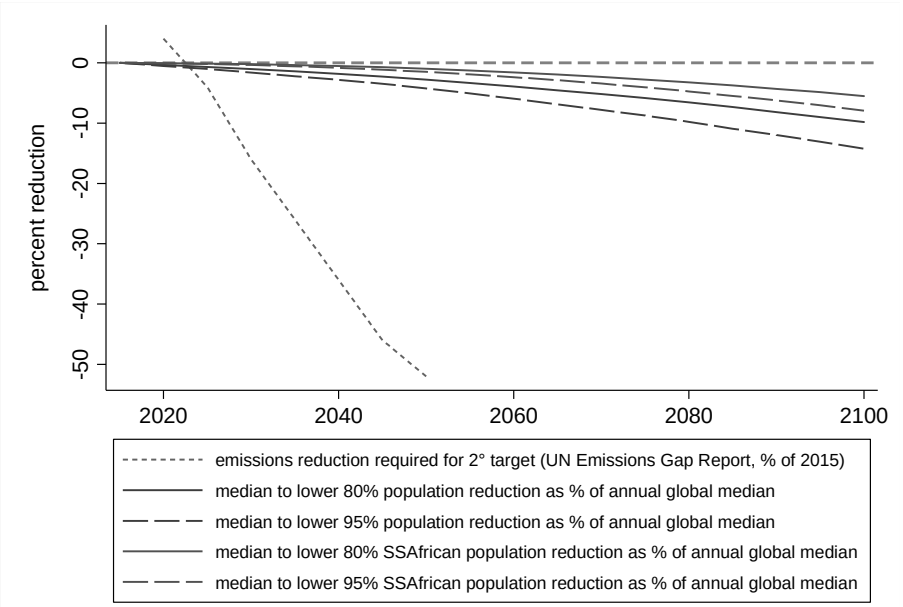
approximately how small the future population could plausibly be, relative to the most likely path. The graph includes two versions of each low path: a version in which the whole world takes the low rather than the medium population path, and a version in which only sub-Saharan Africa take the low path, while the rest of the world remains on the medium path. Although it is not plausible that only sub-Saharan Africa would deviate in this way, this exercise is useful for demonstrating that a large fraction of the uncertainty in the size of the future population comes from uncertainty in the size of the future African population.

Also plotted in Figure 1 is the percent reduction in emissions needed to meet a 2° climate change target, according to the UN Emissions Gap report. The graph compares percent reductions in population with percent reductions in emissions, because O'Neill, et al. (2012) empirically compute an elasticity of approximately one: a one percent reduction in the size of a population would result in an approximately one percent reduction in emissions. However, there is still an important dissimilarity between the emissions line and the population line: the emissions line plots percent reductions *relative to what emissions were in 2015*, while the population lines plot percent reductions *relative to population in that year on the median path*. Even on the lower 95% path, population size is expected to increase over the 21st century and be well above 9 billion in 2100.

Two conclusions are evident in Figure 1. First is that, for the coming decades, the uncertainty in the size of the future population is small. Almost a century into the future, the UN projects that it is very unlikely that the population would be even 10% smaller than the most likely path. Of course, these probabilistic projections are not designed to reflect alternative population policies. Nevertheless, they show that the size of the population is substantially fixed over the coming decades.

The second conclusion from Figure 1 is that the possible variation in the size of the future population is small relative to the needed decline in emissions. The Emissions Gap requirements stop at mid-century, by which time emissions are needed to have fallen dramatically. By 2050, the lower 80% population path is not even 3% below the median path. Any fertility-reduction policy that seeks to meet a large fraction of this emissions gap would have to cause a reduction in the size of the near-term population that would be much larger than even a slowdown in the growth of population size that demographic projections consider very unlikely.

Figure 1. Population uncertainty is small relative to emission decline targets



Note: Authors' computations from UN data. Population projections for illustration are taken from UN World Population Prospects probabilistic projections, which are not intended to represent alternative policy paths. Emissions reduction from UN Emissions Gap Report. SSAfrican = sub-Saharan African.

3.2 Separating Mechanism *i* from Mechanism *ii* in a climate-economy model

Figure 1 presented possibilities for future population size in contrast with needed emissions reductions, but did not compute the consequences of alternative population paths for climate change. That is the task of this section. Here, we use Nordhaus's Dynamic Integrated Climate-Economy (DICE) model, a core part of the work for which Nordhaus shared the 2018 economics Nobel Prize. The DICE model takes an exogenous population path as an input and computes an optimal path over time of carbon taxes, which results, in the model, in an optimal path over time for mitigation policy and decarbonization. DICE can also be used to project economic and temperature consequences under business as usual,[13] where it is assumed that there is no large intensification of mitigation policy.

[13] Here and in what follows, by 'business as usual' we simply refer to the implications of the very low control rates in the "baseline" scenario of Nordhaus's DICE2013, which involve very little mitigation.

Table 1. Peak temperature and optimal taxes in a leading climate-economy model

	future size		business as usual (Mech. *i*)		optimal mitigation (Mech. *ii*)	
population path	2060	2100	2020 tax	peak temperature	2020 tax	peak temperature
median	10.2 b	11.2 b	--	7.1 °C	$25.19	3.1 °C
lower 80%	9.8 b	10.1 b	--	6.7 °C	$23.27	3.2 °C
lower 95%	9.6 b	9.6 b	--	6.5 °C	$22.37	3.2 °C
momentum only	9.0 b	9.0 b	--	6.4 °C	$21.58	3.2 °C

Note: Authors' computations from Nordhaus' DICE 2013 model. "Peak temperature" is the peak temperature increase relative to pre-industrial.

Table 1 presents results for four population paths: the UN's median population projection, the two unlikely low population projections, and a conceptually-minimal population path that reflects only the effectively inevitable consequences of population momentum. The "momentum only" path is the result of a hypothetical exercise in which fertility and mortality rates are immediately brought to replacement levels, so any future population growth is due only to the age structure of the population (as more girls age into their reproductive years).[14] To emphasize, no foreseeable policy implementable by governments could have such an extreme outcomes as the "momentum only" path as a plausible consequence.

The "business as usual" column illustrates Mechanism *i*, the focus of this paper. These estimates ask how much peak temperature would change if only the population path changed, but mitigation policy did not otherwise become more aggressive. In other words, the business as usual results ask how much mitigation in temperature increase could be achieved only by reducing population growth. The result of the computation is that even a hypothetically large, more-extreme-than-achievable, instantaneous change in population growth rates would only, as the core mitigation strategy, reduce peak temperature from a disastrous 7.1 °C to a still-disastrous 6.4 °C. Such a reduction in the temperature change would be an improvement, but not enough of an improvement to constitute a sufficient substitute for other mitigation strategies.

The "optimal mitigation" columns illustrate Mechanism *ii*, which was the focus of Scovronick, *et al.* (2017), although that paper did not consider the population-momentum-only path. In these computations, Nordhaus' DICE model computes

[14] Note that, for countries with negative natural population increase, this exercise could increase fertility rates.

the optimal path of carbon taxes, assuming a given population path. Two facts are noteworthy about these projected peak temperatures. First, they do not vary widely; this is because the model optimizes mitigation policy to avoid too-large temperature change.[15] Second, peak temperature is lower under the lower-growth population path. This is despite the fact that a larger population produces more emissions for a given level of per-capita economic activity and technology. The explanation is that, when the future population will be smaller, fewer people will be harmed by climate change, so the model finds it optimal to mitigate less aggressively. This feature of the model's optimization is visible in the decrease in optimal near-term carbon taxes as future population size decreases. This result is the core of Mechanism *ii*: a smaller future population may not be enough to single-handedly reduce temperature to acceptable levels, but in the context of optimizing policy, it could be a reason to decide to incur smaller mitigation costs.

3.3 Other possible changes in the structure of the population: regional allocation, urbanization, migration

So far, we have only considered possible changes to the total size of the future population, and we have only considered changes within the 21st century. There may be other ways in which a Mechanism *i* effect of the population could operate on emissions, that we do not consider. For example, it may be that the very-long-term sustainability of economic development or survival of the population is improved by the steady state size of the population being 9 billion rather than 12 billion, after such a time as "backstop" technology has replaced carbon emissions, and through a mechanism other than the effect of population size on carbon emissions and therefore peak temperature. This is a longer-term question than 21st century climate policy.

Alternatively, it may also be that the nearer-term composition of the population – its allocation across places – matters for carbon emissions. For example, the age structure of the population, the fraction of the population in the labor force, and the fraction living in rural rather than urban places all predict carbon emissions. At least one such dimension of heterogeneity, however, deepens the challenge: fertility rates are highest (and have the most room to fall) in countries where emissions per capita are low. A fertility policy designed to reduce carbon emissions might reasonably focus on the richest countries, but it would have to overcome the fact that in many

[15] Many readers may consider 3.2°C to be too large of temperature change, and disagree with DICE that this outcome is the optimal balance of mitigation costs and climate damages. We do not take a position on this debate, and only use the DICE model to illustrate the mechanisms that are the focus of this paper.

of these, fertility rates are already historically low and below replacement levels.

In short, although these complications may be important to some questions and may present valuable policy opportunities, it remains quantitatively unlikely that fertility reduction could achieve the level of mitigation called for – in its extent and in its pace – by the Emissions Gap Report.

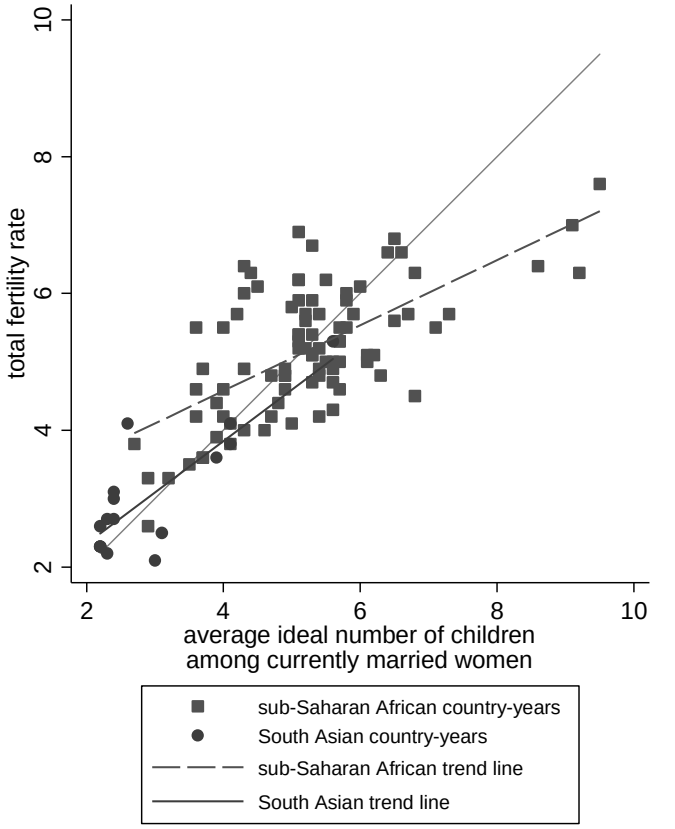
4. Achieved and intended fertility

The "momentum only" projection in Table 1 is merely a theoretical possibility. What would it take to achieve a large decline in fertility rates? In a population where fertility rates are high, an important question is whether women and families *intend* to have high levels of fertility, or whether they would prefer lower levels of fertility but their *achieved* fertility exceeds their intended fertility, perhaps because of lack of access to reproductive health care (Coale and Watkins, 1986). If achieved fertility exceeds intended fertility, then there may be an opportunity for policy to reduce fertility by improving access to contraception. If achieved fertility matches intended fertility, then the policy challenge may be deeper for a program to reduce fertility. In that case, policy would have to induce women to *want* fewer children, a goal that would interact with cultural ideas about fertility as well as economic costs of and opportunities to invest in children's schooling and human capital. Changing parents' perceptions of returns to schooling or ideal family sizes may take decades, even for successful, well-designed programs.

Figure 2 illustrates the issue. It is an update of a graph first drawn by Pritchett (1994). Here, each dot reflects the average outcome of a Demographic and Health Survey, which means that each dot is one developing country in one year. Country-years in South Asia are plotted as circles and country-years in sub-Saharan Africa are plotted as squares (countries in other regions are omitted for clarity). The message of the figure is that survey-reported ideal fertility is highly correlated with achieved total fertility rates. In other words, the countries where women have many children are the countries where women report wanting many children. The correlation is not perfect, but it is not small for empirical cross-country research. Moreover, the slope of the trend line is not far from one-to-one. Taken together, these results suggest that mere prevision of contraception may be unlikely to be enough to quickly and substantially accelerate the decline in fertility rates.[16]

[16] Indeed, it is a common finding in health policy for developing countries that providing hardware is not enough, when behavior change is necessary. Coffey and Spears (2017), for example, document the case where providing latrines is not enough to sufficiently accelerate the decline in open defecation in rural India, because many people do not use them. In the case of open defecation in India, like in Connelly's historical study of population policy, the unfortunately consequence has sometimes turned out to be state-organized or state-permitted coercion (Gupta, et al. 2019).

Figure 2. Survey-reported ideal fertility is highly correlated, across developing country-years, with achieved total fertility rates



Note: Authors' computations from DHS data. Observations are all Demographic and Health Surveys conducted since 2000. For earlier data, see Pritchett (1994) and Lam (2011). The vertical axis, total fertility rate, is a period rate for the three years prior to the survey, based on period age-specific fertility rates of women 15-49. The thin line is the 45° line of equality between achieved and intended fertility.

Another conclusion visible in Figure 2 is a correlation between geography and fertility: fertility rates are higher in sub-Saharan Africa than in South Asia (where fertility in some countries is already at or near replacement levels). As Figure 1 showed, much of the entire world's uncertainty in future population growth is due

Connelly's historical study of population policy, the unfortunately consequence has sometimes turned out to be state-organized or state-permitted coercion (Gupta, et al. 2019).

to uncertainty in the size of the future African population. Frontier empirical and theoretical research in population science debates alternative explanations for why fertility remains high in sub-Saharan Africa. Africa is poorer, but its fertility is higher even when economics is held constant. Tanzania's total fertility rate in 2016, for example, was almost two-children-per-woman larger than India's was in the year when it had the same GDP per capita that Tanzania did in 2016; Nigeria's TFR in 2016 was about three children larger than India's in the year when it had Nigeria's 2016 level of economic wellbeing (Economist, 2018). It would be a further challenge for policy to reduce intended fertility in sub-Saharan Africa if social scientists indeed do not yet fully understand its causes.

5. Conclusion

One of the core questions of population ethics is how policy-makers should weigh changes in average wellbeing against changes in population size. If fertility-reduction policy were a promising tool for reducing carbon emissions, at acceptably low social cost, in the places where emissions are high, on a time scale relevant for climate policy, then population ethicists would have an urgent open task to complete. We would need to know whether such policies were worth it: would it be an improvement, all things considered, to prevent some lives from being lived, so that climate change would be less severe? Because the theoretical questions of population ethics are far from consensus, this would be a worrying need. Fortunately – if not otherwise, at least for this aspect of practical policy-making – we have computed that fertility-reduction policy making is unlikely to be a promising use of scarce political capital and policy attention, as a focal near-term tool of climate mitigation. Of course, this does not mean that human development policy that has the consequence of reducing fertility rates might not be valuable for other reasons.

This also does not mean that the size and the growth of the human population may not be an important input in to climate policy-making for reasons other than Mechanism *i*. However, the practical unimportance of Mechanism *i* partially deflates the argument, common in the philosophical literature, that we must resolve the theoretical questions of population ethics before knowing what to do about climate change. We suspect that, once the empirical facts are correctly registered, any plausible approach to population ethics would conclude that rapid and aggressive decarbonization should be a policy priority. Aggressive mitigation, in technical terms, is a dominating corner solution (Arrhenius, et al, 2019).

As a final note, despite our arguments in this paper, it is good that development professionals and population ethicists are in dialogue in this symposium, because

population ethics is one component of aggregate social welfare. "Policy evaluation" has recently been implicitly redefined, in practice, as near-term, local empirical average impact evaluation. These empirical quantities, such as from a microeconometric experiment, are important to know. But a social objective or axiology is necessary to evaluate whether or not a policy is, all things considered, socially desirable. Although we have argued that the choice of population axiology would not turn out to make an important difference to near-term optimal decarbonization policy, the choice of axiology may make a difference to other potential development policies that have further implications (perhaps including unintended implications) for population growth. So, population ethics could prove to be even more important to development policy than to climate policy.

References

Arrhenius, G., Budolfson, M., and Spears, D., 2019. Deflationary Responses to the Challenges of Population Ethics for Public Policy. In Budolfson, M., McPherson, T., and Plunkett, D. eds. *Philosophy and Climate Change.* Oxford University Press.

Asheim, G.B. and Zuber, S., 2014. Escaping the repugnant conclusion: Rank-discounted utilitarianism with variable population. *Theoretical Economics*, 9(3), pp.629–650.

Barreca, A.I., 2012. Climate change, humidity, and mortality in the United States. *Journal of Environmental Economics and Management*, 63(1), pp. 19–34.

Barreca, A., et al., 2016. Adapting to Climate Change: The Remarkable Decline in the US Temperature-Mortality Relationship over the Twentieth Century. *Journal of Political Economy*. 124(1). pp. 105–159.

Bradshaw, C.J. and Brook, B.W., 2014. Human population reduction is not a quick fix for environmental problems. *Proceedings of the National Academy of Sciences*, 111(46), pp. 16610–16615.

Broome, J., 2012. *Climate matters: Ethics in a warming world*. WW Norton & Company.

Budolfson, M., 2018. Market Failure, the Tragedy of the Commons, and Default Libertarianism in Contemporary Economics and Policy. In Schmidtz, D. and Pavel, C. eds. *The Oxford Handbook of Freedom*. Oxford University Press.

Budolfson, M., Dennig, F., Fleurbaey, M., Scovronick, N., Siebert, A., Spears, D. and Wagner, F., 2018. Optimal climate policy and the future of world economic development. *World Bank Economic Review*.

Budolfson, M. and Spears, D., 2018. Why the Repugnant Conclusion is Inescapable. Princeton University Working Paper.

Caney, S. 2019. People, Planet and Public Policy: The Role of 'Population' in Addressing Climate Change. PPE Conference Presentation.

Coale, A. and Watkins, S.C., 1986. *The decline of fertility in Europe*. Princeton University Press.

Coffey, D. and Spears, D. 2017. *Where India Goes: Abandoned Toilets, Stunted Development, and the Costs of Caste*. HarperCollins: Delhi.

Conly, S., 2016. *One child: Do we have a right to more?*. Oxford: Oxford.

Connelly, M.J., 2009. *Fatal misconception: The struggle to control world population*. Harvard University Press.

Dasgupta, P., 1995. The population problem: Theory and evidence. *Journal of Economic Literature*, 33(4), pp.1879-1902.

The Economist. 2018. Demography: Babies are lovely, but... 22 September, pp. 41–42.

Gerdts, C., et al., 2016).Impact of clinic closures on women obtaining abortion services after implementation of a restrictive law in Texas. *American Journal of Public Health*, 106, pp. 857–864.

Gerland, P., Raftery, A.E., Ševčíková, H., Li, N., Gu, D., Spoorenberg, T., Alkema, L., Fosdick, B.K., Chunn, J., Lalic, N. and Bay, G., 2014. World population stabilization unlikely this century. *Science*, 346(6206), pp. 234–237.

Geruso, M. and Spears, D., 2018. Heat, Humidity, and Infant Mortality in the Developing World. IZA Discussion Paper 11717.

Greaves, H., 2018. Climate change and optimum population. *The Monist*, 102(1), pp. 42–65.

Guillebaud, J., 2016. Voluntary family planning to minimise and mitigate climate change. *BMJ*, 353, p.i2102.

Gupta, A., at al. 2019. Changes in open defecation in rural north India: 2014–2018. IZA Discussion Paper.

Lam, D., 2011. How the world survived the population bomb: Lessons from 50 years of extraordinary demographic history. *Demography*, 48(4), pp. 1231–1262.

Lawson, N. and Spears, D., 2018. Optimal population and exhaustible resource constraints. *Journal of Population Economics*, 31(1), pp. 295–335.

Ng, Y.K., 1989. What Should We Do About Future Generations?: Impossibility of Parfit's Theory *X*. *Economics & Philosophy*, 5(2), pp. 235–253.

Nordhaus, W. and Sztorc, P, 2013, User's Manual for DICE-2013R, online at: http://www.econ.yale.edu/~nordhaus/homepage/documents/DICE_Manual_103113r2.pdf.

O'Neill, B.C., Jiang, L. and Gerland, P., 2015. Plausible reductions in future population growth and implications for the environment. *Proceedings of the National Academy of Sciences*, 112(6), pp. E506–E506.

O'Neill, B.C., Liddle, B., Jiang, L., Smith, K.R., Pachauri, S., Dalton, M. and Fuchs, R., 2012. Demographic change and carbon dioxide emissions. *The Lancet*, 380(9837), pp. 157–164.

Pacala, S. and Socolow, R. 2004. Stabilization Wedges: Solving the Climate Problem for the Next 50 Years with Current Technologies. *Science*. 305, pp. 968-972.

Pritchett, L.H., 1994. Desired fertility and the impact of population policies. *Population and Development Review*, pp. 1–55.

Roberts, M. 2019. Does Climate Change Put Ethics on a Collision Course with Itself? PPE Conference Presentation.

Scovronick, N., Budolfson, M.B., Dennig, F., Fleurbaey, M., Siebert, A., Socolow, R.H., Spears, D. and Wagner, F., 2017. Impact of population growth and population ethics on climate change mitigation policy. *Proceedings of the National Academy of Sciences*, 114(46), pp. 12338–12343.

Sherwood, S.C., and Huber, M., 2010. An adaptability limit to climate change due to heat stress. *Proceedings of the National Academy of Sciences*. 107(21), pp. 9552–9555.

Spears, D., 2015. Smaller human population in 2100 could importantly reduce the risk of climate catastrophe. *Proceedings of the National Academy of Sciences*, *112*(18), pp. E2270–E2270.

Spears, D., 2017. Making people happy or making happy people? Questionnaire-experimental studies of population ethics and policy. *Social Choice and Welfare*, 49(1), pp. 145–169.

Spears, D., 2019. *Air: Pollution, Climate Change, and India's Choice Between Policy or Pretence*. HarperCollins: Delhi.

UN Environment, 2017. *Emissions Gap Report.*

Kirsti M. Jylhä,[1] Pontus Strimling[2] & Jens Rydgren[3]

Climate change denial among radical right-wing supporters[4]

Political right-wing orientation correlates with climate change denial in several Western countries. Politicians and voters of far-right (i.e., radical and extreme right-wing) parties seem to be particularly inclined to dismiss climate change but the reason for this is unclear. Thus, the present paper investigates if and why climate change denial is more common among voters of the radical right-wing party Sweden Democrats as compared to voters of a mainstream right-wing party (the Conservative Party, *Moderaterna*), and compares both these voter groups with left-wing (Social Democrat) voters. In four regression analyses, distrust of public service media (Swedish Television, *SVT*), socioeconomic right-wing attitudes, and negative attitudes toward feminism and women were the strongest predictors of climate change denial. These variables outperformed conservative ideologies (Right-Wing Authoritarianism and Social Dominance Orientation), anti-immigration attitudes, distrust of the Parliament and courts, and belief in conspiracies, in predicting denial. Voter group explained only a small or zero part of variance in denial over and above these variables. The results suggest that even though radical and mainstream right-wing parties emphasize different sociopolitical issues

[1] Institute for Futures Studies, kirsti.jylha@iffs.se.
[2] Institute for Futures Studies & Center for Cultural Evolution, Stockholm University.
[3] Department of Sociology, Stockholm University.
[4] This research was supported by a grant awarded by the Swedish Research Council [grant number 2018-00782] to Kirsti Jylhä, grants awarded by Knut and Alice Wallenberg Foundation [grant number 2016.0167 and 2017.0257] to Pontus Strimling; and a grant awarded by the Swedish Research Council [grant number 2016-01995] to Jens Rydgren. Financial support by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences) is gratefully acknowledged.

and anti-establishment messages, similar psychological factors seem to explain why these voter groups differ from each other and from left-wing voters in climate change denial. However, the included independent variables were intercorrelated, which calls into question to what degree they can be separated when explaining psychological underpinnings of climate change denial.

*

Despite the extensive scientific evidence supporting human induced climate change (Cook et al., 2016), climate change denial still exists in society and contributes to delaying climate action (Cann & Raymond, 2018; Oreskes & Conway, 2010). Being an issue that needs to be solved through wide-ranging political solutions and societal reforms, climate change has become politized in several countries, with politically right-leaning individuals expressing more climate change denial and opposition to climate policies than individuals that lean toward the left (Poortinga, Spence, Whitmarsh, Capstick & Pidgeon, 2011; Hornsey, Harris, Bain & Fielding, 2016; McCright & Dunlap, 2003). Recent analyses suggest that politicians and voters of far-right (i.e., radical and extreme) parties are particularly inclined to dismiss climate change (Lockwood, 2018; Forchtner & Kølvraa, 2015; Forchtner, Kroneder & Wetzel, 2018) but only a few studies have to date empirically investigated possible explanations for this.

Socioeconomic and sociocultural explanations

It has been suggested that protection of the industrialized capitalist system and free-market economy is an important explanation for climate change denial, which could explain why denial is more common among right-wing voters (Hoffarth & Hodson, 2016; McCright, Marquart-Pyatt, Shwom, Brechin & Allen, 2016). Supporting socioeconomic explanations also among *radical* right-wing voters, the correlation between Trump support and climate change denial is partly mediated by aversion to wealth distribution (Panno, Carrus & Leone, 2019). However, many radical right-wing parties tend to take vague positions on socioeconomic issues (Rovny, 2013). Also, their voters come from different parties across the political spectrum and express on average *less* right-leaning socioeconomic preferences than voters of the mainstream right-wing parties (Ivarsflaten, 2005; Jylhä, Rydgren & Strimling, 2019a). Thus, additional explanations need to be explored to increase understanding of why radical right-wing supporters more strongly oppose climate messages.

The sociocultural issues promoted by the radical right could also be considered when explaining their tendency for anti-environmentalism (Jylhä & Hellmer, 2020; Lockwood, 2018). The core issue of the radical right is to limit immigration and they express exclusionary sociocultural preferences in other domains as well, as illustrated in their opposition to multiculturalism and societal focus on minority groups and feminism (Mudde, 2007; Mudde & Rovira Kaltwasser, 2013; Rooduijn, Burgoon, van Elsas & van de Werfhorst, 2017; Rydgren, 2007). In line with this, radical right-wing politicians and voters tend to hold socially conservative and authoritarian ideological attitudes (Mudde, 2007; van Assche, van Hiel, Dhont & Roets, 2018) which strongly predict a generalized tendency to hold negative attitudes towards multiple disadvantaged social groups (Ekehammar, Akrami, Gylje & Zakrisson, 2004; Bergh, Akrami, Sidanius & Sibley, 2016).

Indeed, climate change denial correlates with conservative ideology (authoritarianism and support for group-based hierarchies: Milfont, Richter, Sibley, Wilson & Fischer, 2013; Stanley & Wilson, 2019), negative attitudes toward immigration (Krange, Kaltenborn & Hultman, 2018; Ojala, 2015), and an index capturing different exclusionary sociocultural preferences (opposition to e.g. multiculturalism and feminism: Jylhä & Hellmer, 2020). Also, environment and environmentalism are widely considered as stereotypically feminine, and anti-environmentalism could thus reflect promotion of masculine hegemony (Anshelm & Hultman, 2014; Bloodheart & Swim, 2010). However, these sociocultural views are interrelated and correlate also with socioeconomic attitudes (Bergh et al., 2016; Azevedo, Jost, Rothmund & Sterling, 2019) and it is unknown if they uniquely contribute in explaining variance in climate change denial.

Institutional distrust

Radical right-wing parties tend to accuse societal institutions for promoting internationalization and minority rights at the expense of the (native) people (Mudde, 2007; Mols & Jetten, 2015; Rydgren, 2007). The most important targets of these accusations are the mainstream politicians, with whom the other societal institutions are claimed to conspire. Because of this populist rhetoric, radical right-wing parties both attract distrustful voters and increase political cynicism among their supporters (Rooduijn, van der Brug, de Lange & Parlevliet, 2017).

Institutional distrust correlates also with anti-environmental attitudes and beliefs (Harring & Jagers, 2013; Ojala, 2015; Vainio & Paloniemi, 2011). Overlap between far-right voting, institutional distrust, and climate change denial could be due to a conspiratorial worldview, where politician, scientist, and media are perceived as corrupt and malevolent (cf. Mudde, 2004; Castanho Silva, Vegetti &

Littvay, 2017). Another explanation could be that both climate change denial and the anti-establishment views of the radical right reflect more specifically a distrustful stance toward the liberal and cosmopolitan parts of the establishment, meaning that populist arguments are used more instrumentally to challenge the unwanted processes that these institutions are promoting, and to thereby protect the traditional lifestyles and power structures (Jylhä & Hellmer, 2020; Stavrakakis, Katsambekis, Nikisianis, Kioupkiolis & Siomos, 2017; Rydgren, 2017; see also Lockwood, 2018).

Aims and hypotheses

Only a few studies have empirically investigated why climate change denial is linked to far-right support. To address this gap in the literature, we will run a series of hierarchical regression analyses including simultaneously several variables that have been suggested to explain why right-wing voters in general, or radical right-wing voters in particular, tend to deny climate change (Jylhä & Hellmer, 2020; Lockwood, 2018; McCright et al., 2016; Panno et al., 2019): 1) two indexes for conservative ideology: Right-Wing Authoritarianism (authoritarian submission and aggression, and conventionalism: Altemeyer, 1998) and Social Dominance Orientation (acceptance and promotion of group-based hierarchies: Pratto, Sidanius, Stallworth & Malle, 1994); 2) Socioeconomic attitudes; 3) Exclusionary sociocultural attitudes (negative attitudes toward immigration and feminism), and 4) Institutional distrust and belief in conspiracies. We also investigate if these variables account for the differences between voter groups classified as radical right (Sweden Democrats), mainstream right (Conservative Party, *Moderaterna*) or left (Social Democrats).

We expect that climate change denial is not only predicted by socioeconomic attitudes, but also by sociocultural attitudes, meaning that approving attitudes of free-market economy and societal group-based power structures complement each other in explaining denial. These attitudes were also expected to outperform conservative ideologies in explaining climate change denial in the full model, thereby implicating a possible mediation effect whereby the more proximal right-wing attitudes help explain the correlation between conservative ideology and denial (cf. Jylhä & Hellmer, 2020). Finally, we expected that institutional distrust explains variance in denial over and above the effects of conservative ideology and sociopolitical views (cf. Ojala, 2015), but that these sets of variables are intercorrelated given the liberal and cosmopolitan context of contemporary Sweden.

Method

Participants

Participants were 2217 Sweden Democrat supporters, 634 Conservative Party supporters, and 548 Social Democratic Party supporters, as indicated by the question, 'How would you vote if there were an election for the parliament today'? Age ranged between 18 and 79 among Sweden Democrat voters (M=55.8, SD=15.3) between 18 and 79 among Conservative Party voters (M=55.9, SD=17.0), and between 19 and 79 among Social Democrat voters (M=54.4, SD=17.9). In all voter groups, most respondents were male (72/65/54%) and had either university (37/50/43%) or high school education (50/42/47%).

Data were collected during spring 2018 by the independent research company Novus at the request of the authors. A selection of panelists was invited from the Sweden Panel, a randomly recruited pool of approximately 40,000 volunteers. Also, 239 of the participants were recruited by a market research company Norstats. This study was conducted following the ethical and professional principles from ICC/ESOMAR International Code on Market, Opinion and Social Research and Data Analytics. For full description of data collection, see Jylhä, Rydgren, and Strimling (2019).

Measures and procedure

Climate change denial was measured by item 'Global warming that is caused by humans is happening' (reversed). We also measured *socioeconomic right-wing attitudes* (three items, α = .72, example: 'Taxes should be reduced'), negative *attitudes toward immigration* (three items, α = .94, example: 'Immigration to Sweden should be reduced'), negative *attitudes toward feminism and women* (three items, α = .77, example: 'Feminism has gone too far'), *Right-Wing Authoritarianism* (three items, α = .53, example: 'To stop the radical and immoral currents in the society today there is a need for a strong leader'), *Social Dominance Orientation* (three items, α = .60, example: 'It's probably a good thing that certain groups are at the top and other groups are at the bottom'), *distrust of the Parliament and courts* (two items, α = .83, example: 'To what degree do you trust that *Riksdagen* manages its work?', reversed), *distrust of a public service media* ('To what degree do you trust news reporting from the following media': SVT [Swedish Television], reversed), and *belief in conspiracies* (six items, α = .79, example: 'A lot of important information is withheld from the public due to self-interest of politicians'). Participants indicated their agreement on these items by a scale ranging from 1 (*disagree completely* or *definitely not true*) to 5 (*agree completely* or *definitely true*), or 6 (*don't know*: handled

as missing values) (For full scales, see Supplementary material). We also measured age, gender (female = 0; male = 1), and education level (0 = elementary school or high school; 1 = university education).

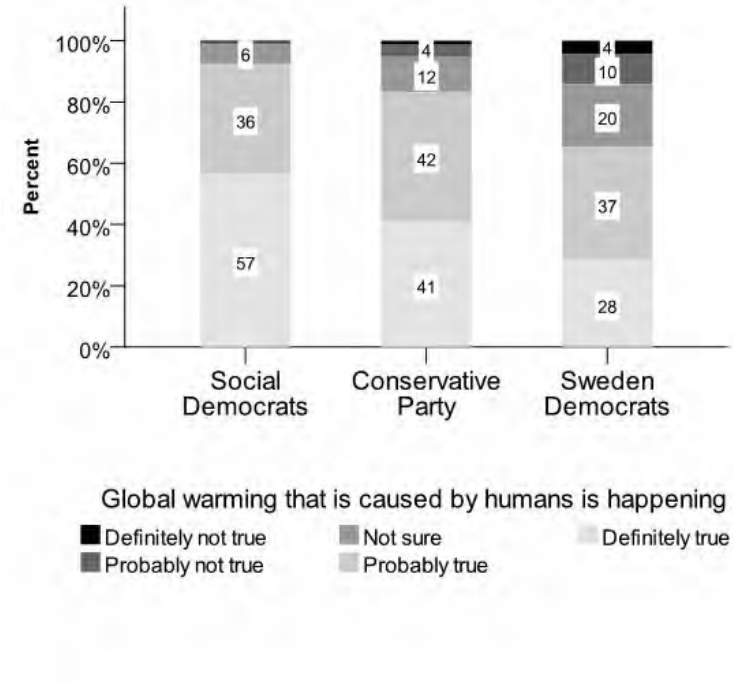
Results

Initial analyses

Majority of respondents agreed that the statement "Global warming that is caused by humans is happening" is *probably* or *definitely true* (65-93%). It was more common to find this statement to be *definitely* or *probably not true* among Sweden Democrat voters (4/10%) than among Conservative Party voters (1/4%) or Social Democratic Voters (0.6/0.7%).

This statement was reverse coded to capture climate change denial. Confirming the above described patterns, Sweden Democrat voters scored highest in climate change denial, followed by voters of the Conservative Party and Social Democrats (see Table 1).

Figure 1. Prevalence of agreeing that human-induced global warming is happening among Social Democrat, Conservative Party, and Sweden Democrat voters



Results of a multivariate ANOVA revealed that, Sweden Democrat voters scored highest in most independent variables, followed by Conservative Party voters and Social Democrat voters (see Table 1), with two exceptions: Sweden Democrat voters scored highest in believing in conspiracies, but Social Democrat and Conservative Party voters did not differ from each other. Conservative Party voters scored highest, and Social Democratic voters scored lowest, in socioeconomic right-wing attitudes.

Table 1. Mean Values (Standard Deviations) and Effect Sizes of Mean Value Differences Between Voter Groups

	Social Democrats	Conservative Party	Sweden Democrats	η^2
	M (SD)	*M (SD)*	*M (SD)*	
Climate change denial	1.53 (0.7)	1.82 (0.9)	2.25 (1.1)	.07
Socioeconomic right-wing attitudes	2.14 (0.8)	3.82 (0.8)	3.58 (0.9)	.28
Anti-feminism	2.28 (1.1)	2.94 (1.0)	3.55 (0.9)	.19
Anti-immigration	2.87 (1.3)	4.07 (1.0)	4.82 (0.4)	.47
Distrust, Parliament and courts	2.41 (0.8)	3.04 (1.0)	3.94 (0.9)	.30
Distrust, Public service media	1.79 (0.7)	2.39 (1.1)	3.37 (1.3)	.22
Social Dominance Orientation	1.57 (0.7)	2.20 (0.8)	2.32 (0.8)	.11
Right-Wing Authoritarianism	1.53 (0.8)	2.97 (0.9)	3.58 (0.8)	.21
Belief in conspiracies	2.42† (0.8)	2.42† (0.8)	2.76 (0.9)	.04†

† = non-significant difference between Social Democrat and Conservative Party voters. All other group differences statistically significant (ps < .01)

Correlation and regression analyses

Climate change denial correlated positively with all independent variables and with the control variables age and gender (see Table 2). Having a university education correlated very weakly with climate change denial (r = –.05) and was thus omitted from the further analyses.

In a series of hierarchical regression analyses predicting climate change denial, independent variables were: conservative ideologies (Step 1), socioeconomic right-wing attitudes (Step 2), exclusionary sociocultural attitudes (negative attitudes toward immigration and feminism) (Step 3), anti-establishment views and belief in conspiracies (Step 4), and party support (Step 5). In each regression analysis, only those participants' data, who supported the parties in comparison, were included.

Across all voter groups, all included sets of psychological variables explained variance in climate change denial (see Table 3). The strongest predictor was distrust of public service media. Socioeconomic attitudes and anti-feminist attitudes explained roughly the same share of variance in denial. Social Dominance Orientation had a weak effect on climate change denial in analyses including Sweden Democrat voters. Party support explained either zero or a very small (1%) part of denial above the effect of these variables, and in one analysis this correlation switched direction from positive to negative indicating a suppression effect due to other intercorrelated variables. The effects of all other variables vanished in the full model. No serious concerns were detected regarding multicollinearity assumptions in analyses including the psychological variables (Tolerances > .52).

Analyses controlling for age and gender did not alter the main results. Age, but not gender, explained some additional variance in climate change (1-2%) among voters of Social Democrats and Conservative Party (β = .15), Social Democrats and Sweden Democrats (β = .12), Conservative Party and Sweden Democrats (β = .13), and Sweden Democrats (β = .13) (ps < .001).

Table 2. Bivariate Correlations Between the Variables

	1	2	3	4	5	6	7	8	9	10
1. Climate change denial										
2. Socioeconomic right-wing attitudes	.30									
3. Anti-feminism	.31	.33								
4. Anti-immigration	.25	.42	.51							
5. Distrust, Parliament and courts	.28	.33	.40	.52						
6. Distrust, Public service media	.36	.40	.38	.41	.58					
7. Social Dominance Orientation	.21	.32	.32	.31	.23	.27				
8. Right-Wing Authoritarianism	.25	.30	.49	.55	.43	.37	.34			
9. Belief in conspiracies	.16	.13	.33	.27	.36	.21	.15	.33		
10. Male gender	.10	.13	.20	.14	.10	.16	.09	.12	–.06	
11. Age	.14	.13	.06	.11	.02†	–0.2†	–.08	.09	.07	.06

† = non-significant, All other correlations statistically significant (ps < .05).

Table 3. Summary of Hierarchical Multiple Regression Analysis Predicting Climate Change Denial in analyses including (1) Mainstream Voters, (2) Mainstream Left-Wing voters and Radical Right-Wing Voters, (3) Right-Wing Voters, and (4) Radical Right-Wing Voters

Variable	1. Social Democrat & Conservative Party		2. Social Democrat & Sweden Democrat		3. Conservative Party & Sweden Democrat		Sweden Democrat	
	ΔR^2	*β*	ΔR^2	*β*	ΔR^2	*β*	ΔR^2	*β*
Step 1	.08***		.08***		.05***		.03***	
Social Dominance Orientation		.17***		.15***		.11***		.11***
Right-Wing Authoritarianism		.18***		.19***		.16***		.11**
Step 2	.04***		.06***		.04***		.04***	
Social Dominance Orientation		.08*		.09***		.08***		.08**
Right-Wing Authoritarianism		.14***		.11***		.15***		.08**
Socioeconomic Attitudes		.22***		.27***		.19***		.21***
Step 3	.02***		.03***		.03***		.03***	
Social Dominance Orientation		.06†		.06**		.06**		.06**
Right-Wing Authoritarianism		.07*		.05*		.07***		.05*
Socioeconomic Attitudes		.19***		.22***		.18***		.20***
Antifeminism		.15***		.19***		.18***		.17***
Anti-immigration		.01		–.01		.04†		–.02

Step 4	.04***		.03***		.04***		.03***	
Social Dominance Orientation		.04		.05**		.05*		.05*
Right-Wing Authoritarianism		.07†		.02		.04†		.02
Socioeconomic Attitudes		.13***		.17***		.14***		.15***
Antifeminism		.13***		.16***		.14***		.14***
Anti-immigration		–.02		–.04		–.00		–.03
Distrust, Parliament and courts		.03		.03		.03		.02
Distrust, public service media		.21***		.18***		.19***		.17***
Belief in conspiracies		–.01		.02		.03		.03
Step 5	.005**		.00		.01***			
Social Dominance Orientation		.05		.05**		.05**		
Right-Wing Authoritarianism		.06†		.02		.03		
Socioeconomic Attitudes		.19***		.16***		.15***		
Antifeminism		.13***		.16***		.14***		
Anti-immigration		–.00		–.04		–.03		
Distrust, Parliament and courts		.04		.02		.02		
Distrust, public service media		.21***		.18***		.18***		
Belief in conspiracies		–.02		.02		.03		
Party support		–.10**		.00		.08***		
Total R^2	.17		.19		.15		.12	
N	1140		2633		2737		2115	

*** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .10$

Discussion

The results showed that majority of participants believe that human-induced climate change is happening. Climate change denial was more common among voters of the radical right-wing party Sweden Democrats than among mainstream right-wing (Conservative Party) voters, and very uncommon among left-wing (Social Democrat) voters. As expected, socioeconomic right-wing attitudes predicted denial (cf., McCright et al., 2016). We found that also anti-feminism has a unique effect on denial, perhaps indicating a link between anti-environmentalism and a motivation to protect the traditional gender norms and masculine hegemony (see Anshelm & Hultman, 2014; Bloodheart & Swim, 2010). The effect of anti-immigration attitudes was weaker than anti-feminist attitudes, possibly because these attitudes were more common and may thus reflect a wide set of underlying psychological motivations. On the other hand, negative attitudes toward women/feminism and immigrants/immigration are strongly correlated (Table 2; see also Bergh et al., 2016) and is it thus questionable if these attitudes can be fully separated in explanations. Dismissal of climate change could be a part of a more general anti-egalitarian worldview where also the uneven distributions of risks and benefits of climate change are more readily accepted (Jylhä, 2016; Jylhä, Cantal, Akrami & Milfont, 2016).

Distrust of public service media was the strongest predictor of climate change denial, which could reflect a doubtful stance toward a media outlet that communicates messages that some voters perceive as undesirable (cf. Schulz, Wirth & Müller, 2018). Distrust of the Parliament and courts did not predict a unique part of variance in denial. Perhaps this variable does not only capture for example cynical perceptions regarding politicians, but also overlaps with the ideological worldviews that a certain sociopolitical system is *not* representing. Indeed, distrust of the Parliament and courts correlated strongly with authoritarian attitudes and negative views on feminism and immigration. The more deeply rooted cynicism regarding politicians' character may not be inherently correlated with climate change denial, as is supported by the weaker correlation between belief in conspiracies and denial (see Table 2: see also Hornsey, Harris & Fielding, 2018) and a recently found weak correlation between anti-political establishment attitudes and denial (Jylhä & Hellmer, 2020). Future studies could investigate more systematically to what degree climate change denial reflects political cynicism or distrust.

Conclusions

Results of a well-powered correlation study showed that, even though mainstream and radical right-wing parties differ in their emphasis on different sociopolitical

issues and anti-establishment messages (Mudde, 2007; Rydgren, 2007; Rovny, 2013), the same variables seem to explain why these voter groups differ from each other and from left-wing voters in climate change denial. The included variables were intercorrelated, and thus it needs to be studied further if – and to what degree – their effects can be separated when explaining climate change denial. Finally, most participants acknowledge human-induced climate change in all voter groups. Thus, although Sweden Democrat voters deny climate change more commonly than voters of the other included parties, denial is not a defining character of these voters as they are clearly more united in their opposition to immigration.

References

Altemeyer, B. (1998). The other "authoritarian personality". In L. Berkowitz (Ed.), *Advance in experimental social psychology* (Vol. 30, pp. 47–92). Orlando, FL: Academic Press.

Anshelm, J. & Hultman, M. (2014). A green fatwā? Climate change as a threat to the masculinity of industrial modernity. *International Journal for Masculinity Studies, 9*, 84–96.

Azevedo, F., Jost, J. T., Rothmund, T. & Sterling, J. (2019). Neoliberal ideology and the justification of inequality in capitalist societies: Why social and economic dimensions of ideology are intertwined. *Journal of Social Issues, 75*, 49–88.

Bergh, R., Akrami, N., Sidanius, J. & Sibley, C. G. (2016). Is group membership necessary for understanding generalized prejudice? A re-evaluation of why prejudices are interrelated. *Journal of Personality and Social Psychology, 111,* 367–395.

Bloodheart, B. & Swim, J. (2010). Equality, harmony, and the environment: An ecofeminist approach to understanding the role of cultural values on the treatment of women and nature. *Ecopsychology, 2*.

Cann, H. W. & Raymond, L. (2018). Does climate denialism still matter? The prevalence of alternative frames in opposition to climate policy. *Environmental Politics*, 1–22.

Castanho Silva, B., Vegetti, F. & Littvay, L. (2017). The elite is up to something: Exploring the relation between populism and belief in conspiracy theories. *Swiss Political Science Review, 23*, 423–443.

Cook, J., Oreskes, N., Doran, P. T., Antilla, W. R. L., Verheggen, B., Maibach, E. W., et al. (2016). Consensus on consensus: A synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters, 11*: 048002.

Ekehammar, B., Akrami, N., Gylje, M. & Zakrisson, I. (2004). What matters most to prejudice: Big five personality, social dominance orientation, or right-wing authoritarianism? *European Journal of Personality, 18,* 463–482.

Harring, N. & Jagers, S. C. (2013). Should we trust in values? Explaining public support for pro-environmental taxes. *Sustainability, 5,* 210–227.

Hoffarth, M. R. & Hodson, G. (2016). Green on the outside, red on the inside: Perceived environmentalist threat as a factor explaining political polarization of climate change. *Journal of Environmental Psychology, 45*, 40–49.

Forchtner, B. & Kølvraa, C. (2015). The nature of nationalism: Populist radical right parties on countryside and climate. *Nature and Culture, 10*, 199–224.

Forchtner, B., Kroneder, A. & Wetzel, D. (2018). Being skeptical? Exploring far-right climate-change communication in Germany. *Environmental Communication, 12*, 589–604.

Hornsey, M. J., Harris, E. A., Bain, P. G. & Fielding, K. S. (2016). Meta-analyses of the determinants and outcomes of belief in climate change. *Nature Climate Change, 6,* 622–626.

Hornsey, M. J., Harris, E. A. & Fielding, K. S. (2018). Relationships among conspiratorial beliefs, conservatism and climate scepticism across nations. *Nature Climate change, 8,* 614–620.

Ivarsflaten, E. (2005). The vulnerable populist right parties: No economic realignment fueling their electoral success. *European Journal of Political Research, 44,* 465–492.

Jacquet, J., Dietrich, M. & Jost, J. T. (2015). The ideological divide and climate change opinion: "top-down" and "bottom-up" approaches. *Frontiers in Psychology, 5,* 1–6.

Jylhä, K. M. (2016). *Ideological roots of climate change denial: Resistance to change, acceptance of inequality, or both?* Doctoral thesis, Uppsala University, Uppsala, Sweden.

Jylhä, K. M., Cantal, C., Akrami, N. & Milfont, T. L. (2016). Denial of anthropogenic climate change: Social dominance orientation helps explain the conservative male effect in Brazil and Sweden. *Personality and Individual differences, 98*, 184–187.

Jylhä, K. M. & Hellmer, K. (2020). Right-wing populism and climate change denial: The roles of exclusionary and anti-egalitarian preferences, conservative ideology, and anti-establishment attitudes. *Analyses of Social Issues and Public Policy*.

Jylhä, K., Rydgren, J. & Strimling, P. (2019a). Radical right-wing voters from right and left: Comparing Sweden Democrat voters who previously voted for the Conservative Party or the Social Democratic Party. *Scandinavian Political Studies.*

Jylhä, K. M., Rydgren, J. & Strimling, P. (2019b). *Sweden Democrat voters: Who are they, where do they come from and where are they headed?* Research report 2019:1. Institute for Future Studies, Stockholm.

Krange, O., Kaltenborn, B. P. & Hultman, M. (2018): Cool dudes in Norway: climate change denial among conservative Norwegian men. *Environmental Sociology*, DOI: 10.1080/23251042.2018.1488516.

Lockwood, M. (2018). Right-wing populism and the climate change agenda: Exploring the linkages. *Environmental Politics, 27*:4, 712–732.

McCright, A. & Dunlap, R. E. (2003). Defeating Kyoto: The conservative movement's impact on US climate change policy. *Social Problems, 50*, 348–373.

McCright, A. M., Marquart-Pyatt, S. T., Shwom, R. L., Brechin, S.R. & Allen, S. (2016). Ideology, capitalism, and climate: Explaining public views about climate change in the United States. *Energy Research and Social Science, 21,* 180–189.

Milfont, T. L., Richter, I., Sibley, C. G., Wilson, M. S. & Fischer, R. (2013). Environmental consequences of the desire to dominate and be superior. *Personality and Social Psychology Bulletin, 39*, 1127–1138.

Mols, F. & Jetten, J. (2016). Explaining the appeal of populist right-wing parties in times of economic prosperity. *Political Psychology, 37,* 275–292.

Mudde, C. (2004). The populist zeitgeist. Government and Opposition, 39, 541–563.

Mudde, C. (2007). *Populist Radical Right Parties in Europe*. Cambridge: Cambridge University Press.

Mudde, C. & Rovira Kaltwasser, C. (2013). Exclusionary vs. inclusionary populism: Comparing contemporary Europe and Latin America. *Government and Opposition, 48*, 147–174.

Ojala, M. (2015). Climate change skepticism among adolescents. *Journal of Youth Studies, 18*, 1135–1153.

Oreskes, N. & Conway, E. M. (2010). *Merchants of Doubt*. New York: Bloomsbury Press.

Panno, A., Carrus, G. & Leone, L. (2019). Attitudes towards Trump policies and climate change: The key roles of aversion to wealth redistribution and political interest. *Journal of Social Issues, 75*, 153–168.

Poortinga, W., Spence, A., Whitmarsh, L., Capstick, S. & Pidgeon, N. F. (2011). Uncertain climate: An investigation into public scepticism about anthropogenic climate change. *Global Environmental Change, 21*, 1015–1024.

Pratto, F., Sidanius, J., Stallworth, L. M. & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology, 72*, 741–763.

Rooduijn, M., Burgoon, B., van Elsas, E. J. & van de Werfhorst, H. G. (2017). Radical distinction: Support for radical left and radical right parties in Europe. *European Union Politics, 18,* 536–559.

Rovny, J. (2013). Where do radical right parties stand? Position blurring in multidimensional competition. *European Political Science Review, 5,* 1–26.

Rydgren, J. (2007). The sociology of the radical right. *Annual Review of Sociology*,

Rydgren, J. (2017). Radical right-wing parties in Europe: What's populism got to do with it? *Journal of Language and Politics, 16,* 485–496.

Schulz, A., Wirth, W. & Müller, P. (2018). We are the people and you are fake news: A social identity approach to populist citizens' false consensus and hostile media perceptions. *Communication Research*, 1–26.

Stanley, S. K. & Wilson, M. S. (2019). Meta-analysing the association between social dominance orientation, authoritarianism, and attitudes on the environment and climate change. *Journal of Environmental Psychology, 61,* 46–56.

Stavrakakis, Y., Katsambekis, G., Nikisianis, N., Kioupkiolis A. & Siomos, T. (2017). Extreme right-wing populism in Europe: Revisiting a reified association. *Critical Discourse Studies, 14,* 420–439.

Vainio, A. & Paloniemi, R. (2011). Does belief matter in climate change action? *Public Understanding of Science, 22,* 382–395.

Appendix: Full scales

Climate change denial

- Global warming that is caused by humans is happening. (R)

Socioeconomic right-wing attitudes

- Taxes should be reduced.
- The public sector is too large.
- It is good to have private profit-driven alternatives in the care sector.

Attitudes toward immigration

- Immigration to Sweden should be reduced.
- Immigration costs too many public resources.
- Immigration leads to increased criminality in Sweden.

Attitudes toward feminism and women

- Feminism has gone too far.
- Women often seek to gain power by controlling men.
- Women tend to interpret harmless remarks or actions as sexist.

Right-Wing Authoritarianism

- To stop the radical and immoral currents in the society today there is a need for a strong leader.
- Our society would be best off if we showed tolerance and understanding for non-traditional values and views. (R)
- The best way to live is in accordance with the old-fashioned values.

Social Dominance Orientation

- It's probably a good thing that certain groups are at the top and other groups are at the bottom.
- We should strive for increased social equality. (R)
- No one group should dominate in society. (R)

Distrust in the Parliament and courts

- To what degree do you trust that the Parliament (Riksdagen) manages its work? (R)
- To what degree do you trust that courts of law manage their work? (R)

Distrust the public service media

- To what degree you trust news reporting the following media? (R)
    - Swedish national public TV. (SVT)

Belief in conspiracies

- A lot of important information is withheld from the public due to self-interest of politicians.
- There is a small, unknown group that really governs world politics and has more power than the elected leaders in different countries.
- There are groups of researchers who manipulate, fabricate or withhold evidence in order to mislead the public.
- The pharmaceutical industry works to keep people sick, rather than healthy, in order to make greater profits.
- Experiments involving new drugs or technologies are conducted on the public without their knowledge or consent.
- Chemtrails, i.e. deliberate discharges of substances from aeroplanes that are used to manipulate people or the weather.

Malcolm Fairbrother,[1] Gustaf Arrhenius,[2] Krister Bykvist[3] & Tim Campbell[4]

How Much Do We Value Future Generations? Climate Change, Debt, and Attitudes towards Policies for Improving Future Lives[5]

Do people care much about future generations? Moral philosophers argue that we should, but it is not clear that laypeople agree. Humanity's thus-far inadequate efforts to address climate change, for example, could be taken as a sign that people are unconcerned about the well-being of future generations. An alternative explanation is that the lack of action is due to public scepticism about climate policies' effectiveness, rather than the discounting of future lives per se. Based on surveys and survey experiments with representative samples of respondents in four countries—Sweden, Spain, South Korea, and China—we find that most people say they care

[1] Institute for Futures Studies, Department of Sociology, Umeå University & Department of Sociology, University of Graz, malcolm.fairbrother@umu.se.
[2] Institute for Futures Studies & Department of Philosophy, Stockholm University, gustaf.arrhenius@iffs.se.
[3] Institute for Futures Studies & Department of Philosophy, Stockholm University, krister.bykvist@iffs.se.
[4] Institute for Futures Studies, tim.campbell@iffs.se.
[5] The authors thank the Swedish Riksbankens Jubileumsfond for the program grant (M17-0372:1) that funded this research. Very helpful comments were provided by participants in seminars at Gothenburg University's Centre for Collective Action Research (CeCAR) and Stockholm University's Institutet för social forskning (SOFI), as well as participants at a conference organized by the Institute for Futures Studies in September 2019.

about future generations, and would even be willing to reduce their standard of living so that people can enjoy better lives in the future. Many do not, however, support policies for reducing either global warming or the national debt—both of which would impose a net cost on current generations for the benefit of future generations. We show that a significant part of the public's apparent lack of concern for future generations is actually due to disbelief or distrust in the likely benefits of government actions.

*

1. Introduction

How can we explain the lack of action in the face of the unfolding climate crisis? Given that scientists have been warning about the problem of climate change for decades, and all that time policy experts have been suggesting ways of responding to it, why has humanity taken so few steps?

One reason for humanity's failure to solve the massive collective action problem that is global climate change could be the fact that climate change is a massively intergenerational issue. Given that the costs of reducing greenhouse gas emissions are incurred immediately while the greatest benefits will be enjoyed in the future, it may be that people alive today simply do not much care about future generations. While moral philosophers and welfare economists ascribe substantial value to future generations (Parfit 1984; see also Arrhenius (forthcoming), (2000); Arrhenius, Ryberg, & Tännsjö (2010); Blackorby, Bossert, & Donaldson (2005); Broome (2004)), perhaps laypeople do not.

This paper investigates the role that the well-being of future generations—both their quality of life and their number—plays in the thinking of current generations with respect to the issue of climate change. Specifically, the paper asks:

(i) How much do people care about future generations? What kinds of people care more versus less? In principle, how willing are people to sacrifice their own standard of living for the benefit of future generations?

(ii) More specifically, how much do people support public policies that would benefit future generations but also entail some sacrifice on the part of current generations?

(iii) To what extent does support for (or opposition to) those policies reflect people's valuation of future generations, versus their beliefs about the policies' effectiveness and/or their trust in major social institutions?

To answer these questions, we report the results of surveys and survey experiments conducted in 2019 in four countries—Sweden, Spain, South Korea, and China. Across these four countries, most people say they care about future generations, and many would even be willing to reduce their own standard of living somewhat if that helped improve people's lives in the future. At the same time, many respondents were unsupportive of two policy actions that government could use to benefit future generations, albeit at some cost to people alive today: reducing either global warming or their country's national debt. We tested how people evaluated policies for reducing either of these two things, for two reasons. First, there are potential linkages between them (as explained further below). Second, while climate change and debt are both issues of intergenerational distribution, policies for mitigating them might appeal to people with rather different political views.

We found that people who report being more concerned about future generations are more supportive of both kinds of policies. So are people who report being more trusting in major social institutions, consistent with a number of prior studies showing that support for environmental policies depends heavily on people's political trust (e.g., Fairbrother 2016a, 2019; Fairbrother et al. 2019; Klenert et al. 2018). Political trust can be defined as positive expectations about the likely behaviours of policymakers and public authorities—the belief that they could but will not do someone trusting them harm—including when they are not being scrutinized (see e.g., Levi and Stoker 2000; Hamm, Smidt, and Mayer 2019). We argue therefore that a lack of concern about the well-being of future generations is not the only reason why a person alive today may fail to support policies intended to benefit future generations. Instead, people may oppose such policies because they do not believe the policies will actually work. Some of our results suggest the latter is in fact the more important reason for people's weak support for future-oriented policies.

This argument speaks to an important debate in scholarship on the ethics and economics of climate change, and climate policy. Though reducing greenhouse gas emissions has a cost, some researchers suggest the cost need not be borne by current generations. By means of public debt, or perhaps a "climate world bank", the costs of climate policies could be deferred to future generations (Broome 2016; Broome and Foley 2016; Sachs 2014). These researchers believe that this would be fair, not only because future generations will be the main beneficiaries of climate policies, but also because future people will probably enjoy higher standards of living (e.g.,

Keramidas et al. 2018).[6] At the same time, some of the same researchers also interpret the current lack of global action on climate change as proof that people today are "just not moral enough" (Broome 2018). There is a certain tension between these two claims: If the costs to present generations of mitigating climate change could be reduced to zero, selfishness cannot explain a lack of action. In contrast, our results point to the prevalence of excessive "effectiveness scepticism", or scepticism about the effectiveness of an environmental policy (Bolderdijk et al. 2017). If that is indeed the major problem, then it could be hard to win people's support even for policies that will cost them little or nothing.[7] It would seem a higher priority to find ways of raising public confidence in the policies' effectiveness rather than looking for creative ways of delaying paying for them.

The remainder of this paper proceeds as follows. First, we contextualize our study by reference to literature on the ethics and economics of climate change, climate policies, and intergenerational fairness. Second, we present the data and research design we employ in our empirical investigation. Third, we present the results from our surveys and survey experiments. Fourth, we conclude with a discussion of the study's limitations and implications.

2. Context and Background

Since emitting greenhouse gases causes harm to others, moral philosophers argue that people should not do it (e.g., Broome (2008), (2012)(1992), Conly 2015). The imposition of costs through the effects carbon pollution are well-known to be directional in time. The externalized costs of greenhouse gas emissions largely flow forward, across generations, making climate change an issue of intergenerational justice—in the sense of being related to "the moral duties owed by present to future people and the rights that future people hold against present people" (Kolstad et al. 2014: 216). The preamble to the United Nations Framework Convention on Climate Change therefore concludes by referring specifically to the signatories' determination "to protect the climate system for present and future generations."

With respect to climate change and many other issues of intergenerational justice, moral philosophers and welfare economists argue we must give weight to the well-being of future generations, and that current generations should be willing to

[6] However, not everyone agrees that shifting the costs to future generations would be fair. For an opposing view, see Gardiner 2017.

[7] To be clear, in focusing on the mass public while seeking to understand humanity's overall failure to address major environmental problems, we are not dismissing the influence of top-down political pressures from elites with a stake in the status quo. Rather, we regard public attitudes as partly a product of such campaigns, and of elite cues. One of the goals of elite campaigns is precisely to shape public views, because the latter's views matter politically (Manza and Brooks 2012).

make some sacrifice on behalf of future generations (see e.g., Parfit 1984; Arrhenius 2000; Arrhenius et al. 2010; Blackorby et al. 2005; Broome 2004). Judging by the public inaction on climate change, however, it seems that the public does not in fact care much about future generations.[8] This lack of concern would make sense given that, as van der Linden et al. (2015) put it: "mounting evidence from across the behavioral sciences has found that most people regard climate change as a nonurgent and psychologically distant risk—spatially, temporally, and socially—which has led to deferred public decision making about mitigation and adaptation responses." Future generations and their well-being may be very far from most people's minds.

The value that people attach to future generations is not well understood, and measuring people's preferences about the temporal distribution of policy benefits is difficult (Jacobs 2016). Few studies have attempted to investigate what people causing the "externalized" costs of climate change—i.e., polluters—think about the future generations whose well-being they are influencing. In 2010, in a rare exception, the International Social Survey Programme asked about people's agreement with the statement "We worry too much about the future of the environment and not enough about prices and jobs today." The distribution of answers on this item was about evenly balanced between agreement and disagreement.

Similarly, some prior studies have looked at discounting—the degree to which people discount the value of well-being in the future (Bernauer 2013; for the morality of discounting, see e.g., Broome (1994); Parfit (1984)). Decisions about climate policy are closely tied to the discount rate applied in cost-benefit analyses, and a fair allocation of climate policy costs and benefits across generations is closely tied to expectations about differences in the standards of living of different generations (Dasgupta 2008; Neumayer 2007). As Neumayer (2007: 301) puts it, "few people would want the future to be worse off than us or would want to violate the inalienable rights of future generations. They are also possibly willing to sacrifice quite a bit for preventing this from happening." In other words, if it is to be fair, the cost burden of mitigating climate change should fall more heavily on people who are richer. Economists' general expectation that future generations will be richer therefore has important implications for what moral philosophers think we should do in terms of climate change (e.g., Broome 2008). Among the lay public,

[8] Of course, another possibility is that the public doesn't believe climate change is real and/or will genuinely affect people's lives. But surveys show that is not actually a widespread view, as Steg (2018) discusses for example with respect to Europe. Recent polls have found more than 70% of Americans believe climate change will harm future generations (Leiserowitz et al. 2019). There are also some people who say climate change is a natural (not significantly anthropogenic) process, and there is little anyone can do to influence it; but such people are few.

likewise, expectations about the incomes of future generations relative to people alive today may strongly influence people's support for policies that will benefit future generations at the expense of current generations. And it is not clear that laypeople share economists' optimistic view that future generations will enjoy higher standards of living than current generations.

The dearth of public actions on climate change—and people's statements that they are not willing to support some future-oriented policy—are not necessarily, however, proof that people are unconcerned about (or discount the well-being of) future generations. Instead, a second possibility is that people are simply unconvinced that some potential measure for mitigating climate change will actually work, or have the benefits ascribed to them. Prior studies have therefore shown, for example, that opposition to environmental taxes is largely driven by people's political distrust (Fairbrother 2016a, 2019; Hammar and Jagers 2006; Harring 2013). Insofar as distrust is a *belief* about the likely behaviours of another—a belief that the behaviours will not be *trustworthy*—we can therefore say that opposition to climate policies can be rooted in either values or beliefs (or both).

This distinction reflects, theoretically, the diversity of ways that the social sciences suggest we can think about the environment. According to one classic and influential perspective in psychology, altruistic attitudes are the very foundation of the environmental movement, including support for environmental policies (Stern 2000; Stern et al. 1999). Such a perspective suggests that supporting environmental protection is, fundamentally, about a willingness to make sacrifices for the benefit of socially, spatially, and/or temporally distant people—plus perhaps non-human species. From this perspective, low public support for key environmental policies, and the inadequacy of humanity's response to major problems like climate change, would seem to be clear evidence of people's selfishness and lack of concern for the well-being of future generations. If previous generations are unwilling to stop imposing costs on future generations, and unwilling to pay any form of compensation for those costs, that is evidence of selfishness (or the opposite of altruism).

But from another perspective, the real costs of even quite aggressive environmental protection are surprisingly modest. Vandyck et al. (2016) estimate for example that mean global temperature increase could be kept at no more than 2˚ for less than a 1% reduction in global GDP. Keramidas et al. (2018) argue that a 2° pathway could be achieved even if global GDP were to more than double between 2020 and 2050. From this second perspective, environmental policy is predominantly an issue not of what people value, but of their beliefs about costs, benefits, and their distribution; environmentalism is not about sacrifice, but social coordination and the improvement of human lives (Fairbrother 2016b). But the complexity of that coordination may make it appear more costly to solve than it

actually is. Jacobs and Matthews (2012) have for example shown, using survey experiments, that people substantially discount the future benefits of public policies, and largely because of uncertainty about the future—including doubts about the likely future benefits of policies.

Uncertainty about whether the state will deliver what it promises undermines support for many policies with long-run benefits (Jacobs 2016). Scepticism about the effectiveness of an environmental policy—effectiveness scepticism—can both lead to opposition, and reflect people's prior dislike of a policy such as because of feelings it is unfair (Bolderdijk et al. 2017). A view of public attitudes as rooted in effectiveness scepticism, and in excessive doubts about the real benefits of public policies, stands in contrast to arguments that people are not moral enough — presumably meaning they do not attach much value to future generations.

As we mentioned earlier, insofar as there are costs associated with addressing climate change, Broome (2016) suggests that it should be possible for intergenerational transfers to be organized such that no generation is disadvantaged. In his argument, current generations would reduce greenhouse gas emissions (at some cost to themselves, and for the sake of future people) but receive de facto compensation for incurring that cost—in the form of consumption paid by debt. Future generations would be burdened with debt, but reap the benefits of reduced climate change. This view reflects an economic take on environmental problems —wherein any such problem is one of injustice, since there is an externalized cost paid by someone other than the polluter (see Fairbrother 2016b). If there are externalities, there is an efficiency loss—and in principle it should be possible to improve efficiency in such a way as to leave nobody worse off. The influential Stern Review of the economics of climate change emphasized how much less it would cost, in total, for humanity to act sooner rather than later to mitigate greenhouse gas emissions (Stern 2007). While doing that would mean current generations paying a price for the benefit of future generations, the overall cost savings to humanity would be substantial—maybe even massively so (Neumayer 2007). If so, though, that means there is an opportunity to reduce the overall cost—it just requires coordination across generations.

To sum up, then, we can distinguish two general (though not completely mutually exclusive) perspectives, which provide potential explanations for what is blocking progress in climate policy. According to the one perspective, the costs of action are large—which means only people willing to pay a significant cost, altruistically, for the sake of others, will support policies. The other perspective takes the costs of action as modest, or even negligible—such that no notable sacrifice is required, but some confidence in the policies/mechanisms is necessary. Each perspective makes a claim about the values people would have to possess in order to support policy

action. Yet no prior study has attempted to assess value-based as opposed to beliefs- or trust-based explanations of the lack of support for key climate policies. This article investigates empirically the degree to which each perspective succeeds in explaining public policy preferences, in four national contexts.

On a final note here, much of the above applies not just to the *quality* of future lives, but also their *quantity*—that is, the impacts of climate change and climate policies on the world's total human population. While it is not an intuitive conclusion for many laypeople, a significant number of moral philosophers and welfare economists argue forcefully that population itself has value (e.g., Broome 2005). That is, ceteris paribus, more human lives are better than fewer, assuming that the additional lives are worth living, or at least if the lives are well worth living—i.e. the good aspects of the life greatly outweigh the bad (e.g., Arrhenius (2000); Blackorby et al. (2005); Broome (2004), (2005); Parfit (1984). Many people may dislike the idea of a growing global population, as they assume a trade-off between quantity of life and quality of life (as we show below to be the case). However, in light of the discussion in moral philosophy and welfare economics about the value of future lives, we investigate public views not only of policies for increasing the quality of future lives, but also the *number* of such lives. Here too we have little prior evidence of public attitudes. The International Social Survey Programme asked nationally representative samples of people in dozens of countries in 2010 to what extent they agreed that "The earth simply cannot continue to support population growth at its present rate." Most people agreed, with relatively modest differences among nations. That question clearly did not ask, though, about whether population growth would be desirable in the *absence* of a trade-off with environmental sustainability and quality of life.

3. Research Design, Data, and Methods

Our empirical investigation proceeds in six stages. First, we describe what people, including people with different demographic characteristics, say about how much they think, care, and are willing to sacrifice for future generations. Second, we present people's self-reported trust in four major social institutions, as preparation for including trust as a predictor in subsequent analyses of relevant outcomes. Third, we present people's support for increasing the world population, including when encouraged to think about a population increase as necessarily implying a lower quality of life. Fourth, we examine people's attitudes towards public policies for reducing either global warming or public debt—framing such policies as a cost to present generations and a benefit to future generations. In particular, we examine the degree to which people's support for such policies correlates with their levels of

concern about future people and their levels of institutional trust. Fifth, we show that people's policy support is closely tied both to their confidence in policies' effectiveness and to their institutional trust. But people's assessments of policies' effectiveness are not only a cause of people's overall policy attitudes; we show they are also a reflection. Sixth, we show that people are more likely to be willing to sacrifice their own standard of living for the sake of future generations if they expect those future generations to be *better* off than themselves. And we further show that people with optimistic outlooks on the future evolution of human standards of living are more trusting, more confident about the benefits of policy interventions (whether climate or debt), more supportive of increasing the population, and more supportive of climate/debt reduction policies.

Sample

Prior studies have shown that public attitudes towards many kinds of policies, including climate and other environmental policies, are heavily conditioned by political trust—including not just an individual survey respondent's political trust, but also that of the whole society in which s/he lives (Fairbrother 2016a). For our empirical study, we therefore conducted surveys, with embedded survey experiments, in four countries with substantially variable levels of political trust: Sweden, Spain, China, and South Korea. Based on prior polls and studies, levels of institutional trust are high in Sweden and China, and low in Spain and South Korea. We also chose these four countries because they span two culturally dissimilar world regions. The surveys were fielded by the international firm Ipsos MORI, using reasonably high-quality, nationally representative samples of adults.[9] Achieved N's were: Sweden 1084, Spain 1298, South Korea 1176, China 1165. Background demographic variables were gender, age, household income, education, and the number of children in the household. The age ranges covered by the samples were: Sweden 16-65, Spain 16-65, South Korea 18-54, China 18-50. The four countries encompass quite varying levels of climate policy performance, with Sweden a strong performer, South Korea a poor performer, and the others in between (Burck et al. 2019).

Survey Questions

Our survey investigated: respondents' self-assessed concern for the well-being of future people; their preferences about the size of the global human population; their

[9] The age ranges covered by the samples varied somewhat: 16-65 in Sweden and Spain, 18-50 in China, and 18-54 in South Korea.

attitudes towards some key public policies; and some of their relevant beliefs and general political views.

We introduced the series of questions we asked respondents by saying: "The next few questions are about how the decisions we make in society today could affect the lives of people who are not even born yet." Note that this statement did not mention climate change, or any specific policy domain. (Depending on the random assignment, some respondents never received a question mentioning global warming.)

To measure people's concerns about future generations, we asked three questions. First, we asked respondents: "How often would you say you think about the lives of future people who have not even been born yet?" Respondents could answer on a five-point scale from ″Never or almost never″ to ″Very often″. Second, we asked: "On a scale from 0 to 10, how much would you say you care or do not care about the future quality of life of people who have not even been born yet? 0 means you do not care at all, 10 means you care a great deal." The purpose of these two questions was to capture people's self-assessed conscientiousness about future generations. Third, as a measure of people's willingness to sacrifice for the sake of future generations, we asked respondents to what extent they would "be willing or not to reduce [their] standard of living, so that people in the future can lead better lives" (on a 0-to-10 scale from not at all willing to completely willing).

Next, after explaining to respondents that "the decisions we make in society today could also influence the size of the world population in the future," we asked people one of twelve versions of a question about being "in favour or not in favour of increasing the population." Respondents could express their opposition or support on 0-to-10 scale. In various different versions, an increased population was said to mean "a lower future standard of living," "no difference to people's standard of living," to be possible even if future people "could definitely enjoy a high standard of living." That randomized treatment was crossed with a randomly assigned reminder either that "increasing the population would mean more people get the chance to live" or "not increasing the population would mean fewer people get the chance to live." The point was to test the impact of different beliefs about future standard of living on preferences about the size of future generations.

Having gotten respondents thinking about the consequences of decisions today for future generations, we then investigated people's support for one of two randomly assigned policy actions that governments could take for the benefit of people in the future. These were framed as "examples" of ways that people today could reduce their standard of living for the sake of improving the lives of people in the future. The two actions were "policies to reduce global warming" and "policies to reduce the national debt." Some respondents, furthermore, received versions of

these questions specifically saying the goal of reducing global warming or reducing the national debt would be achieved "by increasing taxes," (in the case of global warming only) "by paying for more research on new technologies," or (for national debt only) "by cutting spending." Respondents expressed their support on a 0-to-10 scale, from "not support at all" to "completely support." We used the random assignment here to investigate the difference between respondents' views of "policies" generically and specific kinds of policies which experts think would generally be effective but laypeople may not.

Next, we asked about respondents' belief in the policies' effectiveness. On a 0-to-10 scale, from "not confident at all" to "completely confident," we asked respondents how confident they were that the lives of future generations would be improved if the government succeeded in reducing either global warming or the national debt—or if the government said it was introducing certain specific policies towards these ends. By randomly assigning respondents to hear a question either about *actual, achieved reductions* in global warming or the national debt, versus just statements of policies being introduced, we can measure the impact of people's distrust in government claims and/or their intention and ability to achieve what they say they will achieve.

Next, we asked how respondents thought "most people's standards of living will probably change compared to today"—on a five-point scale from "Get much lower" to "Get much higher." We take this as a measure of optimism about the future. And, finally, we also asked about people's trust (on a scale of 0 to 10) in each of a short series of institutions or groups—university research centres, the news media, business and industry, and the national parliament (or congress, in the case of China).

4. Findings

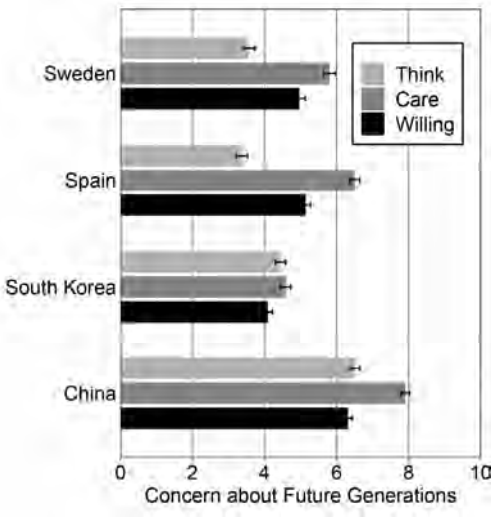
First, we begin by presenting results about people's level of concern about the well-being of future generations—whether people think much and/or care about future people, and would be willing to sacrifice their own standards of living for them. Second, we briefly note what we find about people's answers to the four institutional trust questions, ahead of using trust as a second key predictor (along with concern) of various other attitudes. Third, we consider people's attitudes towards policies for benefiting future generations, comparing the associations between their support for various policies and either concern or trust. Fourth, we compare those results with those for support for increasing the population. Fifth, we examine people's confidence in whether the policies would actually work. And sixth, we consider the issue of people's expectations about future standards of living.

In general, we do not make much of cross-national differences in the average responses to different questions. Comparisons across the four countries must be considered inexact, given that survey questions (translated into different languages) can be received and interpreted differently in different cultural contexts (Davidov et al. 2014). The representativeness and demographic biases of the samples may also differ across the four countries.

(1) Concern about Future Generations

Figure 1 presents the average level of concern people in each of the four countries possess about future generations—judging by respondents' answers to three different questions. (The three questions are about how often respondents think about future people who have not even been born yet; how much they care about the future quality of life of people who have not even been born yet; and about how willing them would be to reduce their standard of living, so that people in the future can lead better lives.) Judging by their answers to these three questions, most people do seem to care at least somewhat about future generations. There was a lot of variation across different people's responses to these questions, but as regards "caring" about future people, for example, a majority of people in all four countries gave an answer of 5 or higher on a 0 to 10 scale (74% in Sweden, 83% in Spain, 54% in South Korea, and 95% in China). Scores for "thinking" about future generations were lower than for "caring", as were those for being willing to sacrifice. In all four countries, a majority of the respondents gave scores of 5 or higher for willingness.

Figure 1. Average concern, by three measures, about future people, by country



The "think" variable, originally measured on a 1–5 scale, has been rescaled to range from 0 to 10.

Responses to the three different questions are correlated in each country (Cronbach's alpha is 0.59 or higher), so we constructed an index of overall concern, using a factor analysis (regression scores, using varimax rotation). We make use of this index in further analyses reported below, but first we can treat it as the outcome in regression models, with age, gender, education, income, and presence of children in the household as predictors—see Table 1.

Table 1: Models of Concern for Future Generations

	Sweden	Spain	S Korea	China
Age	-0.01**	-0.00**	-0.01*	-0.01**
	(0.00)	(0.00)	(0.00)	(0.00)
Male	-0.25**	-0.15**	0.05	-0.07
	(0.06)	(0.05)	(0.06)	(0.05)
Education	0.22**	0.08	0.19**	-0.17*
	(0.06)	(0.05)	(0.06)	(0.08)
Income	-0.03*	0.01	0.01	0.04**
	(0.01)	(0.01)	(0.01)	(0.00)
Child in Household	0.26**	0.20**	0.32**	0.27**
	(0.06)	(0.05)	(0.06)	(0.06)
(Intercept)	0.44**	0.12	-0.08	-0.36**
	(0.11)	(0.11)	(0.11)	(0.13)
N	951	1113	1124	1155

*Coefficients (with standard errors in parentheses). Dependent variable ranges from 0 to 10. **<0.01, *< 0.05.*

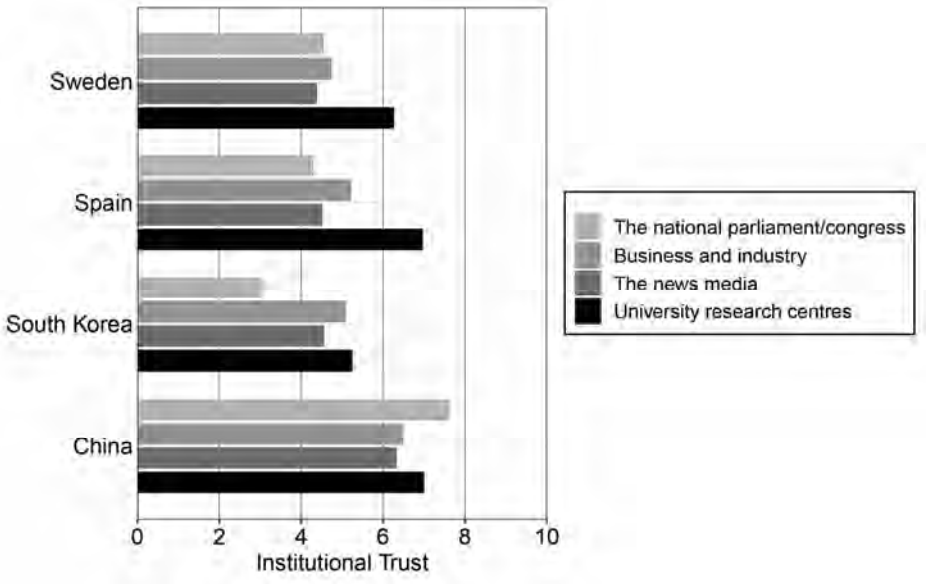
Based on these models, in every country, older respondents expressed less concern about future generations, while respondents in households with children expressed more concern. By comparison, the relationships with gender, education (coded dichotomously as any education beyond secondary or not), and income differed across the four countries.

(2) Institutional Trust

The four questions about trust in major social institutions also correlated with each other—Cronbach's alpha was 0.72 or higher. Levels of institutional trust varied substantially across the four countries—see Figure 2: 47% of respondents in Sweden had an average score of 5 or higher (across the four institutions), 52% in Spain, 87% in China, and 32% in South Korea. We therefore captured the difference we expected in institutional trust between the two Asian countries, but the minimal

difference between the two European countries in their average levels of trust was surprising (as was the fact that the level of institutional trust was slightly higher in Spain than in Sweden).

Figure 2. Average institutional trust, by country

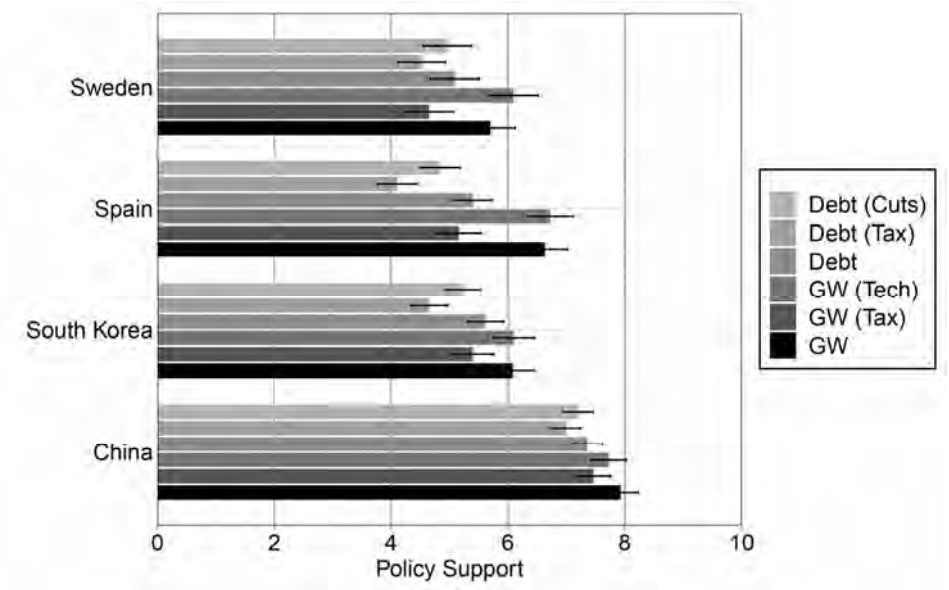


(3) Policy Support

When asked about policies for benefiting future generations, people were moderately supportive—see Figure 3. Support declined if respondents were told the policy entailed paying higher taxes. Respondents were more supportive about helping future generations by reducing climate change than by reducing national debt.[10] But, whatever the issue (climate change or national debt), raising taxes is unpopular. And, otherwise, (randomly assigned) differences among the hypothetical policies do not make much difference.

[10] Note that one of the policies is global (climate change) whereas the other is national (debt). We might have expected less support for climate change, given that many of the benefits of climate policies (i.e., of mitigating greenhouse gas emissions) will accrue to more socially distant people. But that is not what we find.

Figure 3. Average support for different policies in each country



Next, we fit models of policy support—see Table 2. The first model for each country shows only background demographics—age, gender, education (two categories), income, and the presence of a child in the respondent's household. The second model shows coefficients for two randomly assigned treatments—whether the policy was global warming (rather than national debt) and whether it was a tax policy—plus two indices measuring concern about future generations and institutional trust. The third model for each country includes the full set of covariates.

Table 2 shows that background demographics are little related to policy support. In contrast, the four variables included in the second model for each country are all strong predictors of policy support. In such models, all variables are statistically significant, in all countries. Table 2 also shows (observationally rather than experimentally) that support for policies for reducing either climate change or public debt are a function (about equally, pooling all four countries) of both concern and trust.[11] In other words, trust appears to make as much difference to people's policy attitudes as does concern for future generations, generally. And that is true for policies related to either global warming or debt reduction.

[11] We can directly compare the sizes of the coefficients on these two variables, as they are each standardized (centered at zero, and divided by their standard deviations). The beta coefficients here represent the change in Y associated with a 1 standard deviation change in X.

Table 2. Models of Policy Support

	Sweden			Spain			South Korea			China		
Age	-0.03** (0.01)		-0.01* (0.01)	-0.03** (0.01)		-0.02** (0.01)	0.02** (0.01)		0.03** (0.01)	0.01* (0.01)		0.01 (0.01)
Male	-0.04 (0.19)		0.13 (0.16)	-0.30 (0.17)		-0.18 (0.15)	-0.05 (0.14)		-0.07 (0.13)	-0.11 (0.17)		0.08 (0.10)
Education	0.30 (0.19)		-0.26 (0.16)	0.42* (0.18)		0.33* (0.16)	0.12 (0.17)		-0.04 (0.15)	-0.32 (0.18)		-0.15 (0.15)
Income	-0.02 (0.04)		-0.01 (0.03)	0.08** (0.03)		0.05* (0.03)	0.08** (0.03)		0.07** (0.02)	0.03** (0.01)		-0.00 (0.01)
Child in Household	0.41 (0.21)		0.12 (0.18)	0.15 (0.18)		-0.12 (0.16)	0.33* (0.16)		0.01 (0.14)	0.54** (0.13)		0.26* (0.11)
Policy: Glob Warming		0.74** (0.15)	0.65** (0.16)		1.36** (0.14)	1.41** (0.15)		0.81** (0.13)	0.84** (0.13)		0.59** (0.10)	0.58** (0.10)
Policy: Tax		-0.91** (0.16)	-0.82** (0.17)		-1.10** (0.15)	-1.10** (0.16)		-0.67** (0.13)	-0.65** (0.14)		-0.25* (0.10)	-0.25* (0.10)
Concern (Index)		0.98** (0.08)	1.00** (0.08)		0.84** (0.07)	0.86** (0.08)		0.66** (0.07)	0.61** (0.07)		0.43** (0.05)	0.43** (0.05)
Trust (Index)		1.00** (0.08)	1.01** (0.08)		0.53** (0.07)	0.51** (0.08)		0.61** (0.07)	0.62** (0.07)		0.76** (0.05)	0.75** (0.05)
(Intercept)	6.20** (0.37)	5.09** (0.12)	5.71** (0.33)	5.87** (0.35)	5.22** (0.10)	5.62** (0.32)	4.07** (0.30)	5.35** (0.09)	4.12** (0.27)	6.47** (0.29)	7.26** (0.07)	6.93** (0.25)
Valid N	951	1084	951	1113	1298	1113	1124	1176	1124	1155	1165	1155
Adj. R-sq.	0.02	0.31	0.31	0.02	0.23	0.25	0.03	0.21	0.23	0.03	0.29	0.29

*Coefficients and standard errors (in parentheses). Dependent variable ranges from 0 to 10. Significance codes: '**'<0.01, '*'<0.05.*

(4) Support for Increasing the Population

When asked their views about increasing the size of the earth's human population, respondents were lukewarm—see Figure 4. Unsurprisingly, they were less supportive if told the population increase would mean a lower standard of living for future generations, and more supportive if told that the increase would increase or at least not change future generations' standards of living. That people's support increases if they are asked about an increased population and no change in living standards shows that many people, by default, believe that a population increase would affect future people's standards of living. Insofar as respondents suggest they do not want more population, that is partly because they assume more population will mean lower standards of living.

Respondents were also more inclined to support a higher population if reminded it would mean extra people would get to live or that a smaller population would mean fewer people would get to live. That these kinds of manipulations made a difference suggests that without such a prompt people are not fully thinking through the implications of their answers. For this reason, then, we need to be careful about over-interpreting a seeming lack of concern about the size of the population.

Across the four countries, we did not find any consistent demographic correlates of support for rather than opposition to increasing the population.

Figure 4. Average support for increasing population, by country

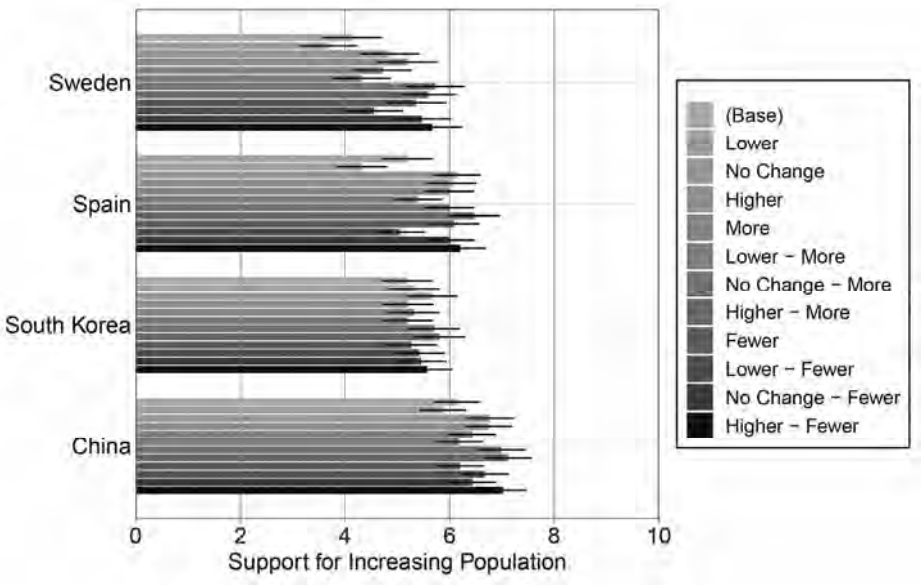


Table 3 presents models parameterizing the relationships represented in Figure 4, plus coefficients for the same two background covariates in Table 2: the three-item index for concern about future people, and the four-item index for institutional trust. Both are strong predictors of support for increased population, as they were of support for policies aimed at future wellbeing. The magnitudes of the relationships are also similar. For both quality and quantity of human life, then, we find evidence that people who are more concerned about future generations and more trusting in major social institutions are more supportive of measures for improving future lives.

Table 3. Models of Support for Increasing Population

		Sweden	Spain	S Korea	China
Change in Living Standards	**Lower**	-0.52*	-0.72**	-0.06	-0.04
		(0.21)	(0.18)	(0.18)	(0.17)
	No Change	0.70**	0.44*	0.14	0.48**
		(0.21)	(0.18)	(0.18)	(0.17)
	Higher	0.74**	0.47**	0.17	0.73**
		(0.21)	(0.18)	(0.18)	(0.17)
Reminder	**More People**	0.68**	0.39*	0.11	0.37*
		(0.18)	(0.15)	(0.16)	(0.15)
	Fewer People	0.78**	0.32*	0.06	0.29
		(0.18)	(0.15)	(0.16)	(0.15)
Indices	**Concern**	0.78**	0.62**	0.61**	0.39**
		(0.08)	(0.06)	(0.07)	(0.06)
	Trust	0.52**	0.56**	0.51**	0.60**
		(0.08)	(0.06)	(0.07)	(0.06)
	(Intercept)	4.21**	5.51**	5.31**	6.04**
		(0.18)	(0.15)	(0.16)	(0.15)
Valid N		1084	1298	1176	1165
Adj. R-sq.		0.19	0.18	0.14	0.16

*Coefficients and standard errors (in parentheses). Dependent variable ranges from 0 to 10. Significance codes: '**'<0.01, '*'<0.05.*

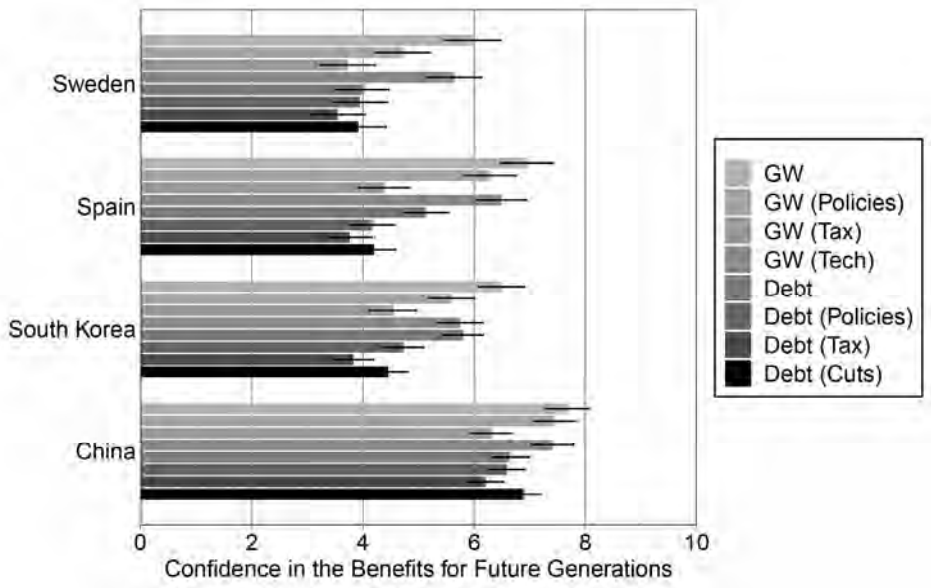
(5) Confidence Versus Effectiveness Scepticism

When asked whether they believed people in the future would really benefit from these policies, respondents' answers were again middling—see Figure 5. Many people appear to be sceptical that policies for reducing global warming or the national debt would actually help future generations. They were significantly less

convinced about the benefits of either cutting spending (to reduce debt) or raising taxes (to reduce either emissions or debt). On the other hand, there was no notable difference between their confidence in the benefits of reducing global warming vis-à-vis cutting the national debt. We also found that trust and confidence are very closely related, much like trust and policy support.

It is difficult to say what causes what: concern about future generations, support for policies, confidence in the policies' effectiveness. We can show, however, that mere mention of taxes changes people's confidence. Respondents who previously received any policy support question about tax (whether for reducing global warming or national debt) were less confident about policies in general. That is, just hearing "tax" made some people less confident, judging by their answers to a subsequent question about government actions generally, including about actions *unrelated to tax*. This shows that effectiveness scepticism is at least to some degree a consequence, not just a cause, of support for or opposition to a policy.

Figure 5. Average confidence in different policies, by country

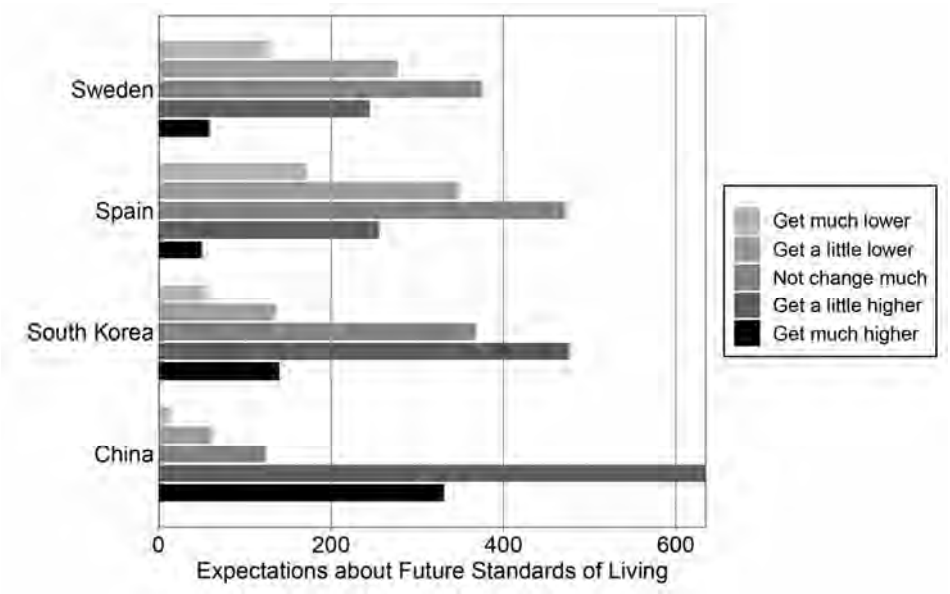


(6) Optimism about Future Standard of Living

Lastly, we asked respondents how they expected standards of living would change in the future. The distribution of the responses appears in Figure 6, and while we

would not want to over-interpret the cross-national differences (given the reasons for caution we articulated earlier), the differences here do some consistent with prior studies about comparative levels of generalized optimism. A YouGov survey in 2015, for example, found Chinese respondents agreed far more than respondents from any other country that the world is "getting better". In contrast, Swedes were far less positive; a majority thought the world is getting worse.[1]

Figure 6. Expectations about future standards of living, by country



In this case, clearly, respondents in the two European countries are far less optimistic about future standards of living. Contrary to what economists generally expect, more of the European respondents said they expected standards of living to decline rather than rise. In the two Asian countries, by comparison, more respondents expected that standards of living would continue rising in the future. We found no demographic variables that consistently predicted more optimism, across the four countries.

There are two possible ways that expectations about future standards of living might be related to people's willingness to sacrifice for the benefit of future

[1] See https://yougov.co.uk/topics/lifestyle/articles-reports/2016/01/05/chinese-people-are-most-optimistic-world.

generations. First, it could be the case that willingness to sacrifice is a consequence of expectations about the future: if so, then people who expect standards of living to decline should be more willing to sacrifice. Alternatively, willingness to sacrifice could reflect general optimism about the future, rooted in positive expectations that sacrifices—and potentially future-oriented policies—will work. In this case, people who expect standards of living to decline should be *less* willing to sacrifice, as they have more negative views of societal functioning, and doubts that any sacrifice they make will in fact benefit future people (perhaps instead of corrupt policymakers and public administrators).

We find the latter view is supported. People who are more optimistic about future standards of living were *more*, not less, willing to sacrifice for future generations—and in all four countries. People who are optimistic about future standards of living are also more trusting, more confident about the benefits of policy interventions (whether climate or debt), more supportive of increasing the population, and more supportive of climate/debt reduction policies. We also found that, in every country, policy support is substantially more correlated with willingness to sacrifice than with the three-item index for concern, or just the other two items on their own. Likewise, in every country, policy support is most correlated with confidence in the policy's effectiveness, which is in turn also more correlated with willingness than with the three-item index for concern.

In sum, then, policy support is more tied to willingness than to the other two items measuring concern (in every country). Willingness appears to be measuring something different than the questions referring to thinking and caring about future generations. Willingness is also more correlated with trust than the other two items. It appears to reflect people's beliefs about the efficacy of sacrificing for the future more than it does people's beliefs about future people's standards of living.

It seems reasonable to think that optimism about future standards of living reflects trust. These two variables correlate (positively), and trust in institutions is a likely reason for people's expectations about the efficacy of their sacrifices. People's support for future-oriented policies reflects their institutional trust more than it does their generalized concerns for future people. We interpret these results to mean there are many people with suspicious outlooks on the world, and their negative views of social institutions—and their pessimism about the effectiveness of key public policies—lead them to be misanthropic.

5. Conclusions

Our empirical study has investigated what value people say they attach to the quality and quantity of future lives, and whether people's apparent lack of concern for

future generations is actually disbelief in the efficacy of policy actions. We have found evidence, based on surveys in four countries, that most people are at least somewhat concerned about future generations. They are even willing to sacrifice their own standard of living, to some degree, so that people in the future can lead better lives. But we have also shown, consistent with prior studies, that many people do not support policy actions that experts say would benefit future generations, at low or even no cost to current generations.

Why does concern about the well-being of future generations not lead to support for policy actions that would contribute to that well-being? Our results suggest that support for such actions is tied not just to the level of people's concern for future generations, but also to their trust in major social institutions, which for many people is not high. Many people do not believe that future-oriented policies will in fact yield significant benefits to people in the future. Many doubt that measures with a short-term cost will actually yield the longer-term benefits that would make them worth the cost. Most people believe that mitigating climate change will make future people's lives better, but they have little confidence that public policies will mitigate climate change. Even if debt could be used to make future generations pay to mitigate climate change, then, current generations might well be suspicious.

Why might doubts about the effectiveness of climate and other future-oriented policies be as prevalent as we have found here? Though this is a topic for another paper, part of the answer may be that measures for environmental protection generally impose significant and concentrated costs on a minority of people: asset-holders and workers in specific industries. We therefore have evidence that workers in polluting industries are therefore less likely to support policies for climate change mitigation (Tvinnereim and Ivarsflaten 2016). And there is now ample evidence that industrial interest groups who stand to lose out from regulatory actions have worked hard politically to prevent or delay those actions (Oreskes and Conway 2010; Farrell 2016; Brulle 2014). One way they have done so is by mounting public campaigns to confuse the broader public, and to spread doubt and misinformation, including about the costs and effectiveness of potential policy responses to the environmental problems their industries cause.

One clear limitation of our study is that we are relying on self-reporting, which may be subject, for example, to social desirability bias. Another potential objection to our study is that we are not adequately quantifying the values we attempt to measure, such as in terms of the metric of money. But, as Neumayer (2007: 300) says, "many effects of climate change simply cannot be adequately monetarily valued."

References

Bolderdijk, Jan Willem, Linda Steg, Edwin Woerdman, René Frieswijk, and Judith I.M. De Groot. 2017. "Understanding Effectiveness Skepticism." *Journal of Public Policy & Marketing* 36[2]: 348–361. DOI: 10.1509/jppm.16.118

Broome, John, and Duncan K. Foley. 2016. "A World Climate Bank." Pp. 156–169 in Iñigo González-Ricoy and Axel Gosseries (eds.) *Institutions For Future Generations*. Oxford: Oxford University Press.

Broome, John. 2005. "Should We Value Population?" *Journal of Political Philosophy* 13[4]: 399–413.

Broome, John. 2008. "The Ethics of Climate Change." *Scientific American* 298: 69–73.

Broome, John. 2016. "Do Not Ask for Morality." Pp. 9–21 in Adrian Walsh, Säde Hormio, and Duncan Purves (eds.) *The Ethical Underpinnings of Climate Economics*. Abingdon: Routledge.

Broome, John. 2018. "Self-interest Against Climate Change." Alf Vanags Memorial Lecture. Riga. May 17th. Viewable at: https://vimeo.com/273639673.

Brulle, Robert J. 2014. "Institutionalizing Delay: Foundation Funding and the Creation of US Climate Change Counter-Movement Organizations." *Climatic Change* 122 (4): 681–94.

Burck, Jan, Ursula Hagen, Niklas Höhne, Leonardo Nascimento, and Christoph Bals. 2019. Climate Change Performance Index: Results 2020. Bonn: Germanwatch. https://www.climate-change-performance-index.org/sites/default/files/documents/ccpi-2020-results-191209.pdf.

Caney, Simon. 2014. "Climate change, intergenerational equity and the social discount rate." *Politics, Philosophy & Economics* 13[4]: 320–342. https://doi.org/10.1177/1470594X14542566.

Dasgupta, Partha. 2008. "Discounting climate change." *Journal of Risk and Uncertainty* 37: 141–169. DOI 10.1007/s11166-008-9049-6.

Davidov, Eldad, Bart Meuleman, Jan Cieciuch, Peter Schmidt, and Jaak Billiet. 2014. "Measurement Equivalence in Cross-National Research." *Annual Review of Sociology* 40: 55–75. doi: 10.1146/annurev-soc-071913-043137.

Fairbrother, Malcolm. 2016a. "Trust and Public Support for Environmental Protection in Diverse National Contexts." *Sociological Science* 3: 359–382.

Fairbrother, Malcolm. 2016b. “Externalities: Why Environmental Sociology Should Bring Them In.” *Environmental Sociology* 2: 375–384.

Fairbrother, Malcolm. 2019. ″When Will People Pay to Pollute? Environmental Taxes, Political Trust and Experimental Evidence from Britain.″ *British Journal of Political Science* 49[2]: 661–682.

Fairbrother, Malcolm, Ingemar Johansson Sevä, and Joakim Kulin. 2019. “Political trust and the relationship between climate change beliefs and support for fossil fuel taxes: Evidence from a survey of 23 European countries.” *Global Environmental Change* 59: 102003.

Farrell, Justin. 2016. ″Corporate funding and ideological polarization about climate change.″ *PNAS* 113 (1): 92–97.

Fleurbaey, Marc and Stephane Zuber. 2013. “Climate Policies Deserve a Negative Discount Rate.” *Chicago Journal of International Law* 13[2]. Available at: https://chicagounbound.uchicago.edu/cjil/vol13/iss2/14.

Gardiner, Stephen. 2017. “The threat of intergenerational extortion: on the temptation to become the climate mafia, masquerading as an intergenerational Robin Hood.” *Canadian Journal of Philosophy* 47: 368–394.

Hamm, Joseph A., Corwin Smidt, and Roger C. Mayer. 2019. ″Understanding the psychological nature and mechanisms of political trust.″ *PLoS ONE* 14(5): e0215835.

Hammar, Henrik, and Sverker C. Jagers. 2006. “Can trust in politicians explain individuals′ support for climate policy? The case of CO2 tax.” *Climate Policy* 5: 613–625.

Harring, Niklas. 2013. ″Understanding the Effects of Corruption and Political Trust on Willingness to Make Economic Sacrifices for Environmental Protection in a Cross- National Perspective.″ *Social Science Quarterly* 94:660-671.

Jacobs, Alan M. 2016. “Policy Making for the Long Term in Advanced Democracies.” *Annual Review of Political Science* 2016. 19:433–54.

Jacobs, Alan M., and J. Scott Matthews. 2012. “Why Do Citizens Discount the Future? Public Opinion and the Timing of Policy Consequences.” *British Journal of Political Science* 42: 903–935. doi:10.1017/S0007123412000117.

Keramidas, K., Tchung-Ming, S., Diaz-Vazquez, A. R., Weitzel, M., Vandyck, T., Després, J., Schmitz, A., Rey Los Santos, L., Wojtowicz, K., Schade, B., Saveyn, B., Soria-Ramirez, A. 2018. "Global Energy and Climate Outlook 2018: Sectoral mitigation options towards a low-emissions economy." Luxembourg: European Union. doi:10.2760/67475.

Klenert, David, Linus Mattauch, Emmanuel Combet, Ottmar Edenhofer, Cameron Hepburn, Ryan Rafaty, and Nicholas Stern. 2018. "Making carbon pricing work for citizens." *Nature Climate Change* 8: 669–677.

Kolstad C., K. Urama, J. Broome, A. Bruvoll, M. Cariño Olvera, D. Fullerton, C. Gollier, W.M. Hanemann, R. Hassan, F. Jotzo, M.R. Khan, L. Meyer, and L. Mundaca. 2014. "Social, Economic and Ethical Concepts and Methods." Pp. 207-282 in *AR5 Climate Change 2014: Mitigation of Climate Change*. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Edenhofer, O., R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, A. Adler, I. Baum, S. Brunner, P. Eickemeier, B. Kriemann, J. Savolainen, S. Schlömer, C. von Stechow, T. Zwickel and J.C. Minx (eds.)]. New York: Cambridge University Press.

Leiserowitz, A., Maibach, E., Rosenthal, S., Kotcher, J., Bergquist, P., Ballew, M., Goldberg, M., & Gustafson, A. 2019. *Climate Change in the American Mind: November 2019*. Yale University and George Mason University. New Haven, CT: Yale Program on Climate Change Communication. https://www.climatechangecommunication.org/wp-content/uploads/2019/12/Climate_Change_American_Mind_November_2019b.pdf

Levi, Margaret, and Laura Stoker. 2000. "Political Trust and Trustworthiness." *Annual Review of Political Science* 3: 475–507.

Manza, Jeff, and Clem Brooks. 2012. "How Sociology Lost Public Opinion: A Genealogy of a Missing Concept in the Study of the Political." *Sociological Theory* 30(2): 89–113.

Neumayer, Eric. 2007. "A missed opportunity: The Stern Review on climate change fails to tackle the issue of non-substitutable loss of natural capital." *Global Environmental Change* 17[3-4]: 297–301.

Oreskes N, Conway EM. 2010. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. New York: Bloomsbury.

Page, Edward A. 2006. *Climate Change, Justice and Future Generations*. Cheltenham, UK: Edward Elgar.

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

Sachs, Jeffrey D. 2014. "Climate Change and Intergenerational Well-Being." Pp. 248–259 in Lucas Bernard and Willi Semmler (eds.) *The Oxford Handbook of the Macroeconomics of Global Warming*. New York: Oxford University Press. https://dx.doi.org/10.1093/oxfordhb/9780199856978.013.0011.

Sander van der Linden, Edward Maibach, Anthony Leiserowitz 2015 "Improving Public Engagement With Climate Change: Five "Best Practice" Insights From Psychological Science", *Perspectives on Psychological Science* 10[6]: 758–763.

Steg, Linda. 2018. "Limiting climate change requires research on climate action." *Nature Climate Change* 8: 754–761. https://doi.org/10.1038/s41558-018-0269-8.

Stern, N. 2007. *The Economics of Climate Change—The Stern Review*. Cambridge: Cambridge University Press.

Stern, Paul C. 2000. "Toward a Coherent Theory of Environmentally Significant Behavior." *Journal of Social Issues* 56(3): 407–424.

Stern, Paul C., Thomas Dietz, Troy Abel, Gregory A. Guagnano, and Linda Kalof. 1999. "A Value-Belief-Norm Theory of Support for Social Movements: The Case of Environmentalism." *Human Ecology Review* 6[2]: 81–97.

Tvinnereim, Endre, and Elisabeth Ivarsflaten. 2016. "Fossil fuels, employment, and support for climate policies." *Energy Policy* 96: 364–371.

Vandyck, Toon, Kimon Keramidas, Bert Saveyn, Alban Kitous, and Zoi Vrontisi. 2016. "A global stocktake of the Paris pledges: Implications for energy systems and economy." *Global Environmental Change* 41: 46–63.

Studies on climate ethics and future generations, vol. 1
Working paper series 2019:1–11

The Bullet-Biting Response to the Non-Identity Problem.
Tim Campbell

Does the Additional Worth-Having Existence Make Things Better?
Melinda A. Roberts

Nondeterminacy and Population Ethics
Anders Herlitz

Can Parfit's Appeal to Incommensurabilities Block the Continuum Argument for the Repugnant Conclusion?
Wlodek Rabinowicz

Positive Egalitarianism
Gustaf Arrhenius & Julia Mosquera

Discounting and Intergenerational Ethics
Marc Fleurbaey & Stéphane Zuber

Population-Adjusted Egalitarianism
Stéphane Zuber

'International Paretianism' and the Question of 'Feasible' Climate Solutions
Katie Steele

Sovereign States in the Greenhouse: Does Jurisdiction Speak against Consumption-Based Emissions Accounting?
Göran Duus-Otterström

On the Alleged Insufficiency of the Polluter Pays Principle
Paul Bowman

Demographic Theory and Population Ethics – Relationships between Population Size and Population Growth
Martin Kolk