WORKING PAPER SERIES

Vol. 1

2019:1–11

STUDIES ON

CLIMATE ETHICS

AND FUTURE GENERATIONS

Working paper series 2019:1–11

Editors: Paul Bowman & Katharina Berndt Rasmussen

Studies on Climate Ethics
and Future Generations
Vol. I

Studies on Climate Ethics and Future Generations Vol. I

Editors: Paul Bowman
Katharina Berndt Rasmussen

Institute for Futures Studies
Working Papers 2019:1-11
Stockholm 2019

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: Annie Spratt
Distribution: The Institute for Futures Studies, 2019

Contents

Preface

The papers in this collection are among the first to be published as part of the Climate Ethics and Future Generations project. The project, which began in 2018 and will run through 2023, is generously financed by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences). The project is led by PI Gustaf Arrhenius and co-PIs Krister Bykvist, and Göran Duus-Otterström.

Climate change is an intergenerational problem. What we – the current generation – do now with respect to climate change will have far-reaching effects on the lives of future generations. It is widely acknowledged that our choices now will affect the *well-being* of those who will exist in the future. It is less widely acknowledged that these choices will also inevitably affect both the number and identities of future persons. That is, our choices concerning what to do about climate change will determine both *who* will exist in the future as well as *how many* people will exist in the future.

The intergenerational aspects of climate change raise urgent questions concerning how we should evaluate choices that will have consequences far into the future, and concerning what our obligations are to future generations. The goal of the Climate Ethics and Future Generations project is to draw on both normative theory and empirical sciences to deliver comprehensive and cutting-edge research on these questions in the context of climate change policy. The project brings together researchers from philosophy, political science, economics, sociology, social psychology, and demography to work on these questions.

The project is organized thematically, in three parts:

I. Foundational value questions in population ethics. The overarching question here is how to evaluate future scenarios in which the number of people, their welfare, and their identities may vary. More specifically, we shall explore whether it is bad if fewer people exist in the future, and whether an increase in population size can compensate for a loss in quality of life.

II. Climate justice. Here we focus on three main questions: Do emissions wrong future people, even if they do not harm them? How should burdens and benefits associated with combating climate change be shared in a just way within and across generations? How, if at all, should future people be represented in democratic decision-making?

III. From theory to practice. This part aims to facilitate getting from theory to practice by focusing on three questions: How should we act when we are not certain about which theories to apply? What are the demographic consequences of climate change? How can we change people's environmental values and attitudes?

The eleven papers in this volume reflect these different themes.

The majority of the papers in the volume (seven of eleven) fall under the first project theme. Each paper considers an important foundational question in population ethics.

In his contribution, Tim Campbell takes up the non-identity problem, which is the problem of explaining why it is wrong to bring into existence a person whose life is seriously and unavoidably flawed when it is both possible and costless to instead bring into existence a person whose life is not similarly flawed. Campbell argues against David Boonin's defense of the so-called "bullet-biting" response to the non-identity problem, which holds that, despite initial appearances, there is no non-identity problem: it is morally permissible to bring into existence the person whose life is unavoidably and seriously flawed. Campbell criticises Boonin's argument for the bullet-biting reply, and the argues for a principle that can ultimately help solve the non-identity problem.

In her contribution, Melinda Roberts considers a principle she calls Pareto plus, which holds that the mere addition of a life worth living makes an outcome better, other things being equal. Although Roberts concedes that Pareto plus is initially plausible, she argues that the arguments in favor of Pareto plus nevertheless fail. According to Roberts, this is a fortunate result given her belief that Pareto plus is potentially dangerous, insofar as the arguably unreasonable demands that it makes on existing and future generations might reduce motivation to make the somewhat more modest sacrifices necessary to address climate change.

In their respective papers, both Anders Herlitz and Wlodek Rabinowicz consider how to avoid the Repugnant Conclusion. This is the conclusion that for some population composed of many people with a high quality of life, there is an outcome that is better that consists of a population composed of a much larger number of people with a positive, but very low quality of life. In his contribution, Herlitz examines Derek Parfit's last views in population ethics, and argues that the combination of Parfit's views enables one to refute common arguments for the Repugnant Conclusion, and he also suggests that the general features of the Parfitian view can provide the basis for a more plausible population axiology more generally.

For his part, Rabinowicz attempts to ground Parfit's view in a wider theory of value (the "fitting-attitudes" theory of value). However, Rabinowicz argues that

accepting this theory of value would require Parfit to give up some of his other substantive value commitments.

In their contribution, Arrhenius and Mosquera note that while most egalitarians have focused on Negative Egalitarianism, which is the view that relations of inequality make things worse, other things being equal, there has been less discussion of Positive Egalitarianism, which is the view that relations of equality make things better, other things being equal. Positive and Negative Egalitarianism diverge, especially in different number cases. Hence, an investigation of Positive Egalitarianism might shed new light on the vexed topic of population ethics and our duties to future generations. In light of some recent criticism, Arrhenius and Mosquera further develop the idea of giving positive value to equal relations.

The next paper is co-authored by two philosophically minded economists, Stéphane Zuber and Marc Fleurbaey. Zuber and Fleurbaey consider the practice of social discounting, which is used by economists to calculate the future costs and benefits of some policy. Roughly, social discounting describes how the value of future costs and benefits changes with time. Zuber and Fleurbaey investigate the ethical issues that social discounting raises, and argue that although social discounting has been mostly discussed in the context of discussions of utilitarian moral theories, social discounting is relevant to a much wider range of moral theories.

Zuber is also the author of the volume's final paper that falls under the first project theme. In his paper, Zuber considers how egalitarian theories (theories that value equality within some population) should be formalized for economic theory in contexts in which population sizes vary. Zuber discusses several different views concerning how economic theory should apply egalitarian criteria in variable population contexts.

The next three papers, by Katie Steele, Göran Duus-Otterström, and Paul Bowman respectively, each address a question central to the second project theme of climate justice. In her contribution, Steele examines the idea of political feasibility in the context of international agreements to address climate change. Specifically, Steele considers the position that the only politically feasible international agreements for addressing climate change are those that make no countries worse off and at least some countries better off. Steele argues that proponents of this position understand political feasibility too narrowly and therefore underestimate the number of agreements that are politically feasible. She argues that acknowledging this fact opens the door for policy-makers to consider a wider range of international agreements on climate change.

Duus-Otterström's paper considers an important question in international

climate policy: whether international bodies should use production or consumption-based accounting of greenhouse gas emissions. Production-based accountting counts emissions at their geographical point of origin, whereas consumption-based accounting counts emissions embodied in some good at the place that the good is consumed. This is an important question of climate justice, since many of the goods that rich, developed countries consume are produced in poorer, developing nations. Ultimately, Duus-Otterström argues that considerations of environmental effectiveness count in favor of the continued use of production-based accounting.

In his contribution, Bowman investigates a common assumption made in the literature on climate justice. This assumption, which he calls the Insufficiency Claim, holds that the polluter pays principle (the principle that those agents who have emitted excessive greenhouse gas emissions should bear the costs of climate change) fails to allocate all of the costs of climate change. Bowman considers why many philosophers have accepted the Insufficiency Claim, and then argues against the claim.

Martin Kolk's paper is the final paper of the volume, and reflects the third project theme. Kolk is a demographer, and in his contribution, he explores the relationship between demographic theory (how population systems regulate themselves, given available resources) and population ethics (how we should value populations). Most theories of population ethics focus on welfare and population size, and demographic theories focus on population size, welfare, and population growth. Kolk considers the interactions among welfare, population growth, and population size.

These papers are among the first to be written for the Climate Ethics and Future Generations Project. We expect that the project will yield significantly more research on these pressing questions. New research and further developments will be continuously updated on www.climateethics.se.

Paul Bowman & Katharina Berndt Rasmussen
Editors

Tim Campbell[1]

The Bullet-Biting Response to the Non-Identity Problem

According to *the bullet-biting response to the non-identity problem*: Given a choice between creating a well-off child, A, and a different child, B, that is significantly worse off than A, it is *not* impermissible to create B. David Boonin has presented an argument for the bullet-biting response. He claims that although the conclusion of his argument is implausible, the rejection of the argument is even more implausible. But Boonin's argument is more implausible than he realizes. Three specific premises, together with the claim that creating a child cannot make that child better or worse off than she would otherwise have been, jointly entail that it is *not* impermissible to create children whose lives contain only pain and suffering. This is a damning objection to Boonin's argument. I argue that this objection cannot be avoided without undermining the other premises of Boonin's argument. Finally, I suggest a fairly weak moral principle that avoids the bullet-biting response. According *The Weak Principle (WP):* If you are choosing between only Act 1 and Act 2, then Act 1 is impermissible if (a) the outcome of Act 2 is significantly better than the outcome of Act 1, (b*) Act 2 wouldn't cause *anyone* to incur a significant cost, and (c) Act 2 wouldn't violate anyone's rights.

[1] Institute for Futures Studies, tim.campbell@iffs.se. Financial support by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences) is gratefully acknowledged.

The Non-Identity Problem is a persisting problem in moral philosophy. It's best introduced by considering examples in which an agent must choose between two different procreative acts. The following example is due to David Boonin:

> *The Non-Identity Case*.[2] Wilma has decided to have a baby. She goes to her doctor and receives some concerning news. If Wilma conceives soon, her child will have a rare health condition—*condition X*—that will cause it to experience chronic pain throughout its life, as well as a significantly shorter-than-average lifespan. The pain will be manageable with treatment. Wilma is willing to, and would certainly, pay for this treatment herself; and she is wealthy enough that these payments wouldn't impose any significant cost on her. With treatment, the child's life would be well worth living, but still significantly worse than the life of a "normal" healthy child.

The good news is that if Wilma takes a pill once a day for two months before conceiving, she will have a perfectly healthy and happy child without condition X. Knowing all this, Wilma decides not to take the pill. As a result, she gives birth to a child with chronic pain and a significantly shorter-than-average lifespan.[3]

Intuitively, Wilma's act is morally impermissible. But it's hard to explain why.[4] According to

> *The default explanation*: Wilma's act is impermissible because it makes her child significantly worse off than it would otherwise have been.

The default explanation is *prima facie* plausible. But there's a well-known objection to it. The sperm and egg that would have joined if Wilma had waited to conceive would have been different from the sperm and egg that join to form the child in the example.[5] For this reason, it seems, the child Wilma would have conceived after waiting *would not have been* the child conceived in the example; it would have been *a different individual*. To emphasize this, Boonin stipulates that the child conceived in the example is a girl named *Pebbles*, and that if Wilma had waited, she would instead have conceived a boy named *Rocks*. If Wilma had waited to conceive, then

[2] Boonin (2014), p. 2. This case is an offshoot of Derek Parfit's often-discussed case *The 14-year-old Girl* introduced in *Reasons and Persons* (1984), p. 358.
[3] Boonin (2014), p. 2.
[4] The non-identity problem isn't always presented as a puzzle about permissibility. Some find it difficult to explain why, for example, Wilma *ought not* to have the worse-off child. Parfit (1984, p. 358) initially presented the non-identity problem as a puzzle concerning how to justify the claim that the 14-year-old girl in his example "ought to wait" and conceive a better-off child later.
[5] See Parfit (1984), pp. 351-355.

Pebbles—whose life is well worth living—wouldn't have existed. Let's assume Pebbles is Wilma's only child. Then the phrase 'her child' in the statement of the default explanation naturally refers only to Pebbles.[6] It's therefore difficult to see how, by conceiving now rather than waiting, Wilma makes *her child* worse off. Yet, it still seems that Wilma's act is impermissible.

If we reject *the default explanation*, we seem to have only two reasonable options: (1) insist that Wilma's act is impermissible even though it doesn't make her child (or anyone else) worse off, and (2) bite the bullet—i.e. claim that Wilma's act isn't impermissible. Recently, Boonin has put forward an intriguing argument for biting the bullet. His central argument is as follows.

> The Non-Identity Argument:
>
> P1. Wilma's act doesn't make Pebbles worse off than Pebbles would otherwise have been.
>
> P2: An act harms someone only if it makes her worse off than she would otherwise have been.
>
> Therefore,
>
> C1. Wilma's act doesn't harm Pebbles. (From P1 and P2)
>
> P3: Wilma's act doesn't harm anyone other than Pebbles.
>
> Therefore,
>
> C2: Wilma's act doesn't harm anyone. (From C1 and P3)
>
> P4: If an act doesn't harm anyone, then it doesn't wrong anyone.
>
> Therefore,
>
> C3: Wilma's act doesn't wrong anyone. (From C2 and P4)
>
> P5: If an act doesn't wrong anyone, then it isn't morally impermissible.
>
> Therefore,

[6] See Parfit (1984), p. 359 for further discussion.

> C4: Wilma's act isn't morally impermissible. (From C3 and P5)[7]

Boonin calls C4 *The Implausible Conclusion*.

To avoid the Implausible Conclusion, one must provide a rationale, something that is at least potentially a reason, for rejecting the Non-Identity Argument. But according to Boonin, not just any rationale will do. He thinks there are three requirements that must be satisfied if one is to provide a *successful* refutation of the Non-Identity Argument. They are:

> (1) *The Independence Requirement*: A claim or set of claims C successfully refutes the Non-Identity Argument only if (i) C refutes some particular premise of the Non-Identity Argument, and (ii) C is supported by reasons other than the reason that accepting C enables one to avoid the Implausible Conclusion.[8]

> (2) *The Robustness Requirement*: A claim or set of claims C successfully refutes some particular premise of the Non-Identity Argument only if C also refutes any weakened version of that premise which, in conjunction with the Non-Identity Argument's other premises, is strong enough to generate the Implausible Conclusion.[9]

> (3) *The Modesty Requirement*: A claim or set of claims C successfully refutes some particular premise of the Non-Identity Argument only if C doesn't have implycations that are more implausible than the Implausible Conclusion.[10]

Are (1)—(3) reasonable requirements of success for refuting the Non-Identity Argument? Should we accept the Implausible Conclusion *if* we accept (1)—(3) as requirements of success? I will argue for a negative answer to both questions. Part I of the paper addresses the first question, Part II, the second.

Part 1: Requirements

In this part of the paper, I argue that Boonin's proposed requirements are highly problematic, that they should be abandoned, and that a new set of requirements

[7] The argument is presented in Boonin (2014), pp. 3-5. Boonin uses the term 'morally wrong' rather than 'impermissible'. This is not a substantive difference between my statement of the Non-Identity Argument and Boonin's; however, I find the concept of moral impermissibility somewhat less obscure than the concept of moral wrongness.
[8] Ibid., p. 20.
[9] Ibid., pp. 21-22.
[10] Ibid., p. 22.

should be sought. I focus mainly on the Independence Requirement (IR) and the Modesty Requirement (MR).

1.1 Independence

IR's clause (i) requires that a successful refutation of the Non-Identity Argument refute some particular premise of that argument. There's reason to reject this requirement. Suppose one argues that torturing a certain baby for fun isn't impermissible. Call this *the baby-torturing argument*, and the problem of how to respond to it *the baby-torturing problem*. Suppose we have the intuition that torturing the baby is impermissible, and so we appeal to

> ***The no baby-torturing principle***: torturing babies for fun is impermissible.

One might reply that our attempt to refute the baby-torturing argument is unsuccessful because the no-baby-torturing principle lacks independent support; the only potential reason to accept it is that it avoids the conclusion that it's okay to torture babies.

This reply isn't entirely convincing. If an argument's conclusion is *very counter-intuitive*, this might provide a reason to provisionally reject the argument, even if it's unclear exactly which of the argument's premises should be rejected.

The foregoing is consistent with the possibility that, upon further reflection, we would find the baby-torturing argument to be *less implausible* than the no baby-torturing principle. Perhaps there are weighty theoretical considerations that support the argument's premises, and perhaps the intuition in support of the no-baby-torturing principle can be debunked or blunted. In that case, there might be *good reason* to reject the no-baby-torturing principle. But then this reason wouldn't be provided by the fact that the principle was extrapolated from a single intuition; it would be provided by countervailing theoretical evidence that either outweighs or undercuts the evidence provided by that intuition.

We have considered a problem for IR's clause (i), but clause (ii) is problematic as well. Boonin gives the following example of a principle that he claims violates (ii):

> Q*: If in either of two outcomes the same number of people would ever live, it would be [impermissible] to select the outcome in which those who live are worse off, or have a lower quality of life, than those who would have lived, at least if the

> outcome would be significantly worse than the alternative and other morally relevant considerations are equal.[11]

According to Boonin, there is no rationale for accepting Q* "other than the fact that [this] would enable us to avoid the Implausible Conclusion."[12]

Importantly, although it is not clear in Boonin's initial characterization of IR, he thinks that the reason referred to in the statement of IR is *contrastive*—it is a reason for accepting C *rather than* some alternative claim that one is considering (where this other claim is not simply the negation of C).[13]

But this renders IR implausible as a condition for successfully refuting the Non-Identity Argument. For example, suppose that two theories, T1 and T2, have all the same implications except that T1 implies the Implausible Conclusion while T2 doesn't. In this case, it seems T2 is more plausible, and hence, more acceptable than T1. (For the sake of the example I'm bracketing considerations, such as simplicity, which plausibly bear on what makes a theory plausible or acceptable.) But it seems our only reason for accepting T2 rather than T1 would be that T2 (but not T1) avoids the Implausible Conclusion. And so given IR, we could not appeal to the more plausible theory, T2, to refute T1. But this methodological restriction seems unprincipled.

1.2 Modesty

According to The Modesty Requirement (MR) any claim that successfully refutes the Non-Identity Argument must not have implications that are more implausible than the Implausible Conclusion. One potential difficulty with this requirement is that there is a non-trivial possibility that every *fully general* population ethics theory entails something that is more implausible than the Implausible Conclusion.[14] (If this were true, MR would be problematic in interesting ways that I lack the space to explore here.)

But, setting this issue aside, MR is problematic. To see why, we should first acknowledge that whatever Boonin's conditions for being a successful refutation of the Non-Identity Argument are, he ought to recognize parallel conditions for the success of the Non-Identity Argument itself. Otherwise, there would be a self-

[11] Boonin (2014), p. 178. My formulation of Q* is slightly different from Boonin's. For example, instead of an "other things being equal" clause, Boonin's formulation contains specific clauses, e.g. "as long as bringing about the better outcome wouldn't violate anyone's rights".
[12] Ibid., p. 179.
[13] That Boonin intends this "contrastive" interpretation of IR comes out in his discussion of challenges to P5 in Chapter 6, Section 6.1.2, pp. 157-169.
[14] Arrhenius (2015), for example, has proven several impossibility theorems demonstrating that we must reject at least one of several extremely intuitively plausible claims.

serving asymmetry between the rules for evaluating claims Boonin accepts and the rules for evaluating claims that challenge those Boonin accepts. To avoid this, the Independence and Modesty Requirements should be replaced with the following general conditions:

> **The General Independence Requirement (GIR):** A claim or set of claims C successfully refutes an argument A only if there's some reason for accepting C other than that accepting C enables one to avoid the conclusion of A.

> **The General Modesty Requirement (GMR)**: A claim or set of claims C successfully refutes an argument A only if C doesn't have implications that are more implausible than A's conclusion. Moreover, A is successful only if A doesn't have implications that are more implausible than A's conclusion.

These general requirements avoid the double-standard mentioned above.

However, given GMR, the Non-Identity Argument seems unsuccessful. This is because the Non-Identity Argument seems to have implications that are more implausible than the Implausible Conclusion. For example, according to P2: An act harms someone only if it makes her worse off than she would otherwise have been. A standard objection to P2 is that creating an utterly miserable person, someone whose life contains only what is bad for her, harms her, but it doesn't make her worse off than she would otherwise have been. If the miserable person hadn't been created, then she wouldn't have *been*; hence, she wouldn't *otherwise* have been. And hence, she isn't worse off than she would otherwise have been.[15]

If creating a miserable person doesn't make her worse off than she would otherwise have been, then the conjunction of P2, P4, and P5 has an implication that's more implausible than the Implausible Conclusion, namely

> *The Very Implausible Conclusion*: Creating an utterly miserable person isn't impermissible.

P2 and the claim that creating a miserable person doesn't make her worse off than she would have been jointly entail:

> (a) Creating an utterly miserable person doesn't harm her.

[15] For an early discussion of this case, see McMahan (1981).

P4 and (a) jointly entail

> (b) Creating an utterly miserable person doesn't wrong her.

Finally, P5 and (b) jointly entail

> (c) Creating an utterly miserable person isn't impermissible. (The Very Implausible Conclusion)

Since The Very Implausible Conclusion is more implausible than the Implausible Conclusion, and follows from the Non-Identity Argument's premises P2, P4, and P5 (and the observation that creating a person doesn't make her worse off than she would otherwise have been), the General Modesty Requirement implies that the Non-Identity Argument is unsuccessful.

1.3 Robustness to the rescue?

One possible reply to the foregoing challenge would be to appeal to the Robustness Requirement (RR). According to RR, a claim or set of claims C successfully refutes a premise of the Non-Identity Argument only if C refutes any weakened version of the premise that, in conjunction with the Non-Identity Argument's other premises, is strong enough to generate the Implausible Conclusion. Even if some of the Non-Identity Argument's premises (independently or jointly) violate the General Modesty Requirement, the Implausible Conclusion might still be established by appealing to weaker versions of these premises.

However, it seems likely that any weakened version of P2, P4, or P5 that's strong enough to establish the Implausible Conclusion (in conjunction with the Non-Identity Argument's other premises) will violate either the General Modesty Requirement (GMR) or the General Independence Requirement (GIR). To see this, first consider P2. To avoid the Very Implausible Conclusion, we could replace P2 with a principle that clearly implies that creating the miserable person harms her. For example, we might appeal to

> P2*: One harms a person only if one causes this person to be in a state that's overall bad for her.

Since creating the miserable person causes her to be in a state that's overall bad for her, P2* implies that creating this person harms her, and hence, P2* avoids the Very Implausible Conclusion.

But even after replacing P2 with P2*, we still violate the General Modesty Requirement. P2* implies, for example, that killing your child, and hence, causing your child to be in a state that is (plausibly) neither overall bad nor overall good, but *neutral* for it, wouldn't harm it. (If you disagree with this assumption about the neutrality of being dead, take the state of being in a dreamless coma instead.) This implication and the other premises of the Non-Identity Argument jointly entail

> *The Crazy Conclusion*: killing your otherwise happy child wouldn't be impermissible.

P2* and the assumption that killing your child would cause your child to become dead, and that an individual's being dead entails that it is in a state that is neither good nor bad for it jointly entail

> (d) Killing your child wouldn't harm your child.

P4 and (d) jointly entail

> (e) Killing your child wouldn't wrong your child.

P5 and (e) jointly entail

> (f) Killing your child isn't impermissible (The Crazy Conclusion).

A more promising substitution for P2 is

> P2**: An act harms a person only if it makes her worse off than she would otherwise have been OR causes her to exist in a state that's overall bad for her.

The first disjunct of P2** avoids the Crazy Conclusion. Killing your child would certainly make her worse off than she would otherwise have been. Moreover, the second disjunct of P2** avoids the Very Implausible Conclusion, since it entails that creating the miserable person harms her; thus P2** seems to satisfy the General Modesty Requirement (GMR).

However, P2** seems to violate the General Independence Requirement (GIR). There's no apparent reason to accept P2** (rather than Boonin's original premise, P2) other than that P2** (but not P2) avoids the Very Implausible Conclusion. (Both

avoid the Crazy Conclusion.) And this problem seems to arise for substitutions of P4 and P5 as well. In each case, it seems the most promising way to establish the Implausible Conclusion while avoiding the Very Implausible Conclusion is to invoke a special exception to the original premise—one that gives the intuitively correct result for cases involving the creation of miserable people. Thus, we end up with things like

> P4*: An act wrongs someone only if it harms them OR causes them to be in a state that's overall bad for them.

or

> P5*: An act is impermissible only if it wrongs someone OR causes them to be in a state that's overall bad for them.

These weakened versions of the original premises violate the General Independence Requirement.

I conclude that Boonin's requirements are unreasonable as stipulative conditions for any successful refutation of the Non-Identity Argument. A different set of conditions is needed.

Part 2: Rejecting the implausible conclusion

In this part of the paper, I argue we should reject P5 of the Non-Identity Argument even if we accept Boonin's three conditions. Thus, for example, even if the challenges put forth in Part I can be answered, we should reject the Implausible Conclusion.

2.1 The moderate principle

Boonin suggests that perhaps the most promising strategy for refuting P5 appeals to

> ***The Moderate Principle*** **(MP)**: If you are choosing between [only] Act 1 and Act 2, then Act 1 is impermissible if (a) the outcome of Act 2 is significantly better than that of Act 1, (b) Act 2 wouldn't cause you to incur a significant cost, and (c) Act 2 wouldn't violate anyone's rights.[16]

The term 'cost' in clause (b) refers to a shortfall in well-being. I say more about how to interpret 'significantly better' below.

[16] Boonin (2014), p. 152.

MP seems to satisfy the Robustness and Independence Requirements, but is also not as strong as a general theory like act-consequentialism, which, Boonin claims (and many others would probably agree), violates the Modesty Requirement.

However, Boonin argues that MP fails to refute the Non-Identity Argument. He distinguishes two versions of MP, and argues that the version needed to refute P5 violates both the Modesty Requirement (MR) and the Independence Requirement (IR). The two versions of MP differ only with respect to what they assume is the class of individuals whose well-being matters for determining outcome-value. (For simplicity, I assume here that nothing other than well-being contributes to outcome value.) According to what Boonin calls the Inclusive Version of MP (or IMP), the value of an outcome is determined by aggregating well-being across everyone in that outcome. I will assume that on this version of MP, Act 2 produces a "significantly better" outcome than Act 1 if and only if the total of well-being for those who exist in the outcome of Act 2 is "significantly greater" than the total of well-being for those who exist in the outcome of Act 1, and that the difference in well-being between Pebbles and Rocks (the two possible children in the Non-Identity Case) is sufficiently large that the outcome in which Rocks is conceived has significantly more well-being than the outcome in which Pebbles is conceived. I will assume that any well-being difference larger than this one also counts as "significant".

In contrast with IMP, according to what Boonin calls the Exclusive Version of MP (or EMP), the value of an outcome is determined by aggregating well-being across all and only those people who are actual. On this version of MP, Act 2 produces a significantly better outcome than Act 1 if and only if Act 2 produces significantly more well-being than Act 1 *for the actual people*.[17]

Boonin claims there is no rationale for accepting IMP (rather than EMP) except that this enables one to avoid the Implausible Conclusion and that therefore IMP violates the Independence Requirement (IR). I think Boonin is mistaken about this. There are many shortcomings of views, such as EMP, that take only actual people's well-being to determine the value of outcomes, but these are *not* reasons to worry about IMP. For example, a rationale for accepting IMP instead of EMP is that only the former is compatible with the plausible claim that, other things being equal, an outcome is better if it contains *more well-being*. This implies, for example, that if outcome A is exactly like outcome B except that A has an additional person who is extremely well-off, then A is better than B. IMP is compatible with this claim, but EMP isn't; according to EMP, if the people in B are the actual people, then the well-being of the extra person in A (who is *not* an actual person) makes no contribution to the value of any outcome, and so on EMP, A and B are equally good.

[17] For discussion of versions of MP, see Boonin (2014), Chapter 6, Section 6.1.1.

2.2 Modesty and the weak principle

In this section, I consider Boonin's claim that IMP violates the Modesty Requirement (MR). I argue that if this is true, we can appeal to an alternative principle that satisfies MR.

Boonin raises three objections to IMP. I will focus only on two of these—I call them the Replacement Objection and the Procreation Objection. This section focuses on the Replacement Objection. The next focuses on the Procreation Objection. For simplicity, I omit the discussion of one of Boonin's objections, though the principle I introduce in this section avoids this objection in the same way that it avoids the Replacement Objection.

The Replacement Objection is based on the following case.

> *The Rescue Case*: Greg faces a choice between two possible acts: Act 1: Saving the life of a young child (a stranger) who is trapped in a burning building. Act 2: Conceiving a child with his partner.

Acts 1 and 2 are mutually exclusive for the following reason: If Greg rescues the trapped child, he will be badly burned and become infertile, and will be unable to conceive children.

The welfare level of the child conceived under Act 2 is the same as that of Rocks—the better-off of the two possible children in the Non-Identity Case. The welfare level of the child rescued under Act 1 is the same as that of Pebbles. Therefore, for the same reason that, in the Non-Identity Case, the outcome of creating Rocks has significantly more well-being than the outcome of creating Pebbles, in the Rescue Case the outcome of Act 2 (conceiving) has significantly more well-being than the outcome of Act 1 (rescuing). Act 2 (conceiving) doesn't impose a significant cost on Greg or violate anyone's rights. Knowing this, Greg chooses Act 1 (rescuing).[18]

Greg's act of rescuing the trapped child seems not only permissible but supererogatory. However, IMP has the counterintuitive implication that it is *impermissible*. The Rescue Case satisfies all of IMP's conditions: Conceiving a child (Act 2) wouldn't impose a significant cost on Greg, wouldn't violate anyone's rights, and would produce significantly more well-being, and hence, a significantly better outcome, than would rescuing the trapped child (Act 1).

The Replacement Objection is powerful and may persuade some to reject IMP. However, there's a weaker principle that avoids this objection, namely

[18] Boonin (2014), p. 274.

> *The Weak Principle (WP):* If you are choosing between [only] Act 1 and Act 2, then Act 1 is impermissible if (a) the outcome of Act 2 is significantly better than the outcome of Act 1, (b*) Act 2 wouldn't cause *anyone* to incur a significant cost, and (c) Act 2 wouldn't violate anyone's rights.

The only difference between IMP and WP is that WP's second clause, (b*), refers to costs incurred not only by *you* (the agent) but by *anyone*. WP doesn't apply to the Replacement Case because that case doesn't satisfy condition (b*). If Greg had conceived (Act 2) rather than rescued (Act 1), the trapped child would have incurred a significant cost—she would have died in the fire.

Unlike the inclusive version of the Moderate Principle (IMP), WP avoids the Replacement Objection. But like IMP, WP avoids the Implausible Conclusion. WP entails that in the Non-Identity Case, Wilma's conceiving Pebbles (Act 1) is impermissible. The Non-Identity Case satisfies all of WP's conditions: the outcome of Wilma's conceiving Rocks (Act 2) is significantly better than the outcome of Wilma's conceiving Pebbles (Act 1). Conceiving Rocks (Act 2) wouldn't result in a significant cost to anyone; and it wouldn't violate anyone's rights.

2.3 Procreation

Unfortunately, WP cannot avoid *the Procreation Objection*. To illustrate the objection, consider

> *The Procreation Case.* Jane is considering having a child. Due to the significant amount of well-being this child would have, the outcome of having it would be significantly better than the outcome of not having any child. Having a child wouldn't impose a significant cost on anyone or violate anyone's rights. Jane decides not to conceive.

The Weak Principle entails it's impermissible for Jane not to conceive. Call this *The Harsh Conclusion*. Some will find this conclusion counterintuitive, and to avoid it they may wish to further qualify our restricted consequentialist principle, WP.

But I think we can bite the bullet. Even if the Weak Principle (WP) entails the Harsh Conclusion, it doesn't violate the Modesty Requirement (MR). WP violates MR only if it has some implication that is more implausible than the Implausible Conclusion; and I don't think the Harsh Conclusion is more implausible than the Implausible Conclusion.

One way of garnering support for the claim that the Harsh Conclusion is less implausible than the Implausible Conclusion is to consider scaled-up versions of the

Non-Identity and Procreation Cases. Suppose Wilma can either create billions of people at Rocks's welfare level or create the *same number* of (different) people at Pebbles's welfare level (they would be happy overall, but have chronic pain and a significantly curtailed lifespan), and suppose Wilma chooses the latter option. Creating the billions at the higher welfare level wouldn't have imposed a significant cost on anyone or violated anyone's rights.

Next, suppose Jane can either create billions of people at Rocks's welfare level or create no one, and she does the latter. Creating the billions wouldn't have imposed a significant cost on anyone or violated anyone's rights.

Now consider the following pair of claims:

> (1) Wilma's act of creating the billions at Pebbles's welfare level (rather than creating the billions at Rocks's welfare level) *isn't impermissible*.
>
> (2) Jane's act of creating no one (rather than creating the billions at Rocks's welfare level) *is impermissible*.

I am strongly compelled to judge that (1) is *more implausible* than (2). And this compels me to judge that the Implausible Conclusion is more implausible than the Harsh Conclusion. For the only difference between claims (1) and (2), on the one hand, and the Implausible Conclusion and the Harsh Conclusion, on the other, concerns the number of individuals that the claims in each pair refer to. But if this is the only difference, then it seems one who accepts a certain implausibility ordering for the claims about the scaled-up cases should accept the analogous implausibility ordering for the claims about the scaled-down cases.

Conclusion

Boonin has given an argument, the Non-Identity Argument, for biting the bullet in response to the non-identity problem. Boonin's defense of his argument depends on a set of conditions for determining the success of any refutation of it. I argued that Boonin's conditions are unreasonable; they are inherently problematic and make it harder to defend the Non-Identity Argument, since, once they are made fully general, they seem to undermine this argument. I also argued that if we accept these conditions, we should reject the Implausible Conclusion by rejecting P5. Boonin thinks perhaps the most promising way to reject P5 is offered by the inclusive version of the Moderate Principle (IMP), but that this principle violates the Independence and Modesty Requirements. I argued that IMP doesn't violate the

Independence Requirement and even if it violates the Modesty Requirement we can appeal to the Weak Principle (WP), which appears to satisfy this requirement.

References

Arrhenius, Gustaf. (2015) Population Ethics: The Challenge of Future Generations. (Unpublished Manuscript).

Boonin, David. (2014) The Non-Identity Problem and the Ethics of Future People. New York: OUP.

McMahan, Jeff. (1981) "Problems of Population Theory." Ethics 92: 96–127.

Parfit, Derek. (1984) Reasons and Persons. Oxford: OUP.

Melinda A. Roberts[1]

Does the Additional Worth-Having Existence Make Things Better?

Let's call the principle that says that the mere addition of the worth-having existence (other things equal) makes things morally better *Pareto plus*. If we accept Pareto plus, then it seems we should also say that some additions that make at least some person at least a little worse off also may – depending on the numbers – make things morally better. I find that latter claim potentially dangerous. As a main focus of an argument why we ought to do something about climate change, I think it moves people not to do anything much about climate change. People won't accept that they or their progeny – or the many, many future people beyond their own progeny – should bear a cost, perhaps a *significant* cost, just to bring ever more future people into existence. But we can't just reject Pareto plus out of hand. For a number of arguments seem to compel us to accept Pareto plus. My goal in this paper is to identify some of the most interesting of those arguments and to show how, in each case, the argument on closer analysis fails.

1 College of New Jersey, robertsm@tcnj.edu. Financial support by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences) is gratefully acknowledged.

Let's call the principle that says that the mere addition of the worth-having existence (other things equal) makes things better *Pareto plus*. Why is it important to test – to retest – Pareto plus? Why is Pareto plus so dangerous?

If we accept Pareto plus, then it seems we should also accept the idea that there is at least some addition that isn't *mere* addition – some addition that makes at least some person worse off; perhaps just a *little* worse off – that also makes things better.

Graph 1: Two Cases: Mere Addition and Not *Mere* Addition

	c1	c2		c3	c4
probability	1	1		1	1
wellbeing	f1	f2		f3	f4
+10	**p1, p2. . . pn**	**p1, p2. . . pn, q**		**p1, p2. . . pn**	**p2. . . pn, q**
+9					**p1**
+0	*q**			*q**	

c1 at f1 and c2 at f2 exhaust alternative choices and alternative accessible futures in one case, c3 at f3 and c4 at f4 in the other; boldface indicates people who do or will exist; italics with the asterisk people who will never exist. Wellbeing is raw, unadjusted value, which value in turn determines (I will say) when a future x is better for a person p than a future y is for a person q.[2] *The term person includes not just many human beings but many non-human animals as well.*

Consider Graph 1. If f2 is better than f1, then so is f4 better than f3. If adding q to f2 adds value to f2 and makes f2 better than f1, then it's hard to deny that adding q to f4 adds value to f4 and that, the cost to p1 being *very* small, it adds *enough* value to make f4 better than f3.

If both Pareto plus and this *little* extension of Pareto plus are true, then . . . I have nothing more to say. But suppose they aren't and we fail to recognize that fact. And suppose we apply the theory we construct on the basis of that mistake to practical issues, including climate change. Then I am concerned that we will leave our audience unmoved to do anything about climate change, indeed moved to do nothing about climate change, and all on the basis of a mistake. For our audience just isn't going to accept that they or, more importantly, their progeny, or, more importantly still, the many, many future people beyond their own progeny, should incur a cost, perhaps a *significant* cost, just to bring ever more future people into existence.

2 I would thus reject the idea that the "person affecting restriction" means that "trade-offs across people" are "ruled out." Bader (unpublished ms.).

Thus the hope that it's a *person-based theory* that will turn out to be true, a theory that will let us say in the first case that f1 is exactly as good as f2 and in the second case that f3 is better than f4.[3] Those implications are I think implications our audience – that is, people asked to sacrifice so that something can be done about climate change – can accept.

At the same time, I recognize that we can't just reject Pareto plus out of hand. For the fact is that a number of arguments seem to compel us to accept Pareto plus. My goal is to identify the most interesting of those arguments and determine for each whether it's as compelling as it seems or has been claimed to be.

I will sketch just a few of those argument and offer replies to each in this paper.

I. Argument from symmetry

Consider Graph 2. The *procreative asymmetry* states that f1 is worse than f2 and c1 is wrong but that f3 isn't worse than f4 and c3 is permissible. More intuitively: it makes things worse to bring the miserable child Meg into existence but doesn't make things better to bring the happy child Holly into existence.

Graph 2: Two Cases: Miserable Child and Happy Child

	c1	c2		c3	c4
probability	1	1		1	1
wellbeing	f1	f2		f3	f4
+10					**Holly**
+0		*Meg**		*Holly**	
-10	**Meg**				

The *argument from symmetry* challenges that intuition. From Peter Singer: "[W]e do seem to do something bad [and make things *worse*] if we knowingly bring a miserable being into existence, and if this is so, it is difficult to explain why we do not do something good [why we do not make things *better*] when we knowingly bring a happy being into existence." But that last point is just Pareto plus.[4]

3 Consider a revised case in which the number of well off people is very great and the cost to the one person is still very small. Depending on the details of the case, the result that the addition of those people makes the future worse will violate certain conditions many find plausible (e.g. Arrhenius 2011). EVNPBI doesn't imply that f3 is worse than f4 in the original case and wouldn't generate the parallel result in the revised case. But in virtue of the fact that its necessary condition would be satisfied in all such cases, it does leave the door open for those results (which I don't find implausible) – that is, for other person-based principles to generate the result that the one future is worse than the other.
4 Singer (2011), p. 108.

Reply: *The loss distinction thesis*. First, a definition:

> A person p sustains a *loss* in a given future if and only if there is an accessible alternate future that is better for p (an accessible alternate future in which p has more wellbeing).

And a concession:

> *Nonexistence comparability*. A future in which a person never exists can be better (or worse) for that person than an alternate future in which a person exists.

Applied here, the view is that Meg sustains a loss at f1 since f2, where she never exists, is better for her than f1, and Holly sustains a loss at f3, even though she doesn't exist there, since f4 is better for her than f3.

Then, the thesis itself:

> *Loss distinction thesis*. For each person p (miserable or happy; existing, future or never existing at all), a loss p sustains at a future x has (far-reaching, cross-world) moral significance if and only if p does or will exist at x.

> (Note: If a loss is morally significant, its moral significance isn't arbitrarily *contained* to any one future or choice; morally significant losses have far-reaching cross-world moral significance. Thus a morally significant loss bears not just on our evaluation of the future in which it is sustained and the choices at that future that give rise to that loss but also on the evaluation of farther flung futures and choices. See Part II below.)

According to the loss distinction thesis, Meg's loss at f1 has full, cross-world moral significance and thus counts against f1 as compared against f2 and against c1. Holly's loss at f3 has no moral significance and doesn't count against f3 as compared against f4 or against c3.

Thus a distinction between the two cases, and an account of why it is that Meg's existence makes things worse but Holly's existence doesn't make things better.

The loss distinction thesis may sound a little deontic – or, heaven forbid, commonsensical. But it overlaps and is consistent with a particular way of formulating the *person based intuition*.

> *Expansive very narrow person based intuition (EVNPBI)*: c at x is wrong and a future x is worse than a future y only if there is a p and an alternate accessible future z such that:
>
> p does or will exist in x *and*
> x is worse for p than z.

This principle is *very narrow* since, where p does or will exist in x, it provides only a necessary, not a sufficient, condition on when x is worse than y and c is wrong, and a condition only on x's being worse, not better, than y and on c's being wrong, not c's being permissible.

In another respect, however, this principle is *expansive*. To determine whether the necessary condition on x being worse than y is satisfied, we can't consider only how p fares in x as compared against *y* but must expand the sweep of the analysis to consider as well how p fares in x as compared against *z*.

Applied to the asymmetry, EVNPBI implies that f3 isn't worse than f4 and that c3 isn't wrong. And it's consistent with the claim that f1 is worse than f2 and that c1 is wrong. Thus it leaves room for other person based principles to step in and say that f1 is worse than f2 and that c1 is wrong.

II. Argument from moral status of merely possible people

Many cases – including Addition Plus (Graph 3) – show that the claim that the merely possible, relative to a given future, e.g., the actual future, have moral status. If that's so, and if we accept nonexistence comparability, then it might seem we are compelled to accept Pareto plus.

Graph 3: Addition Plus

	c1	c2	c3
probability	1	1	1
wellbeing	f1	f2	f3
+11		**Ann**	
+10	**Ann**		
+5			**Ann, Ben**
+1		**Ben**	
+0	*Ben**		

Reply: The loss distinction thesis helps here as well. It recognizes that merely possible people matter morally just as much as you and I do – that we all have the same moral status. But it also says that having moral status just means that, for each of us, *some* of our losses have moral significance and *others* don't.

EVNPBI, in contrast to two variations on *moral actualism*, is again consistent and overlapping with the account of Addition Plus that the Loss Distinction Thesis provides.

III. Argument from the Mere Addition Principle (MAP)

Let's call the principle that says that the addition of the worth having existence doesn't (other things equal) make things worse the *Mere Addition Principle*, or *MAP*. It might seem that we could accept the seemingly innocuous MAP but reject Pareto plus. In fact, however, we can rework a line of reasoning from John Broome (specifically, his argument against the *neutrality intuition*) to show that that's not the case.

We start with Broome's own example:

Graph 4: Three Outcome Case (Broome's case)

	c1	c2	c3
probability	1	1	1
wellbeing	f1	f2	f3
+10			**Jamie**
+5	**p1 ... pn**	**p1 ... pn, Jamie**	**p1 ... pn**
+0	*Jamie**		

Then, the argument:

1. MAP (assumption)
2. EVNPBI (assumption)
3. f2 is at least as good as f1 (from (1))
4. f3 is at least as good as f1 (from (1))
5. f1 is at least as good as f2 (from (2))
6. f1 is at least as good as f3 (from (2))
7. f1 is exactly as good as f2 (conceptual truth, (3), (5))

8. f1 is exactly as good as f3 (conceptual truth, (4), (6))
9. f2 is exactly as good as f3 (conceptual truth, (7), (8))
10. f3 is better than f2 (same-people Pareto[5])
11. Inconsistency (conceptual truths, (9) and (10))
12. If MAP, then EVNPBI is false. (conditional proof, reductio)
13. If EVNPBI is false, then Pareto plus.
14. If MAP, then Pareto plus.

Line (13) isn't obvious. But we can argue for it as follows: If it's not the case that it doesn't make things worse (other things equal) to leave the happy person out of existence, then it makes things worse (other things equal) to leave the happy person out of existence; i.e. it makes things better (other things equal) to bring that person into existence; i.e. Pareto plus.

Reply: We should accept the argument – and the conclusion (14) – but reject MAP. We should, that is, take the position that f2 is actually worse than f1 (as well as f3). EVNPBI itself implies that f1 is exactly as good as f3; and it leaves the door open for same-people Pareto to say that f2 is worse than f3 and thus worse than f1.

Unlike formulations of the person based intuition that don't require the sweeping inquiry, EVNPBI is consistent with the result that f2 is worse than f1. For its necessary condition for f2 being worse than f1 is satisfied in this case: since f2 is worse for p than *f3* is, we are free to take the position that f2 is after all worse than *f1*.

IV. Argument from consistency (Independence)

Compare the Three Outcome Case against the Two Outcome Case.

Applied to this case, EVNPBI implies that f1 is exactly as good as f2 – which seems right, in this case. But that is inconsistent with what we said before about the Three Outcome Case – it's inconsistent with the position that in that case f2 is worse than f1. To avoid the inconsistency, we should then abandon that position, that is, resurrect MAP and, with MAP, Pareto plus.

5 According to same-people Pareto, when futures f1 and f2 contain exactly the same existing and future people, and f2 is better for at least some of those people and worse for none of those people, f2 is better than f1.

Graph 5: Two Outcome Case

	c1	c2
probability	1	1
wellbeing	f1	f2
+10		
+5	**p1 ... pn**	**p1 ... pn, Jamie**
+0	*Jamie**	

Reply: More precise nomenclature will clarify that f1 and the f2 in the Two Outcome Case are distinct from f1 and the f2 in the Three Outcome Case. Once that is done, the inference to f3 is better than f1 in the Three Outcome Case fails (and we avoid inconsistency or any violation of independence).

The distinction in futures from the one case to the other isn't just coincidental; we can't resurrect the original problem by leaving everything else about the two cases alone and then stipulating that the f1 (f2) in the one case is identical to f1 (f2) in the other case. Why not?

> *Accessibility axiom*: If y is an accessible alternative future relative to x, then necessarily y is an accessible alternative future relative to x.

In other words, if y is accessible to x in any case, then y's existence as such a future is built into the very identity of x, possible futures, or worlds, having all their features necessarily.

V. Argument from simple nonidentity

In this case, c1 and c2 exhaust the agents' choices and f1 and f2 the alternative accessible futures; p and q are (of course) distinct.

Graph 6: Simple Nonidentity

	c1	c2
probability	1	1
wellbeing	f1	f2
+10		**q**
+8	**p**	
+0	*q**	*p**

It's standardly said about this case (by Parfit and many others, complying with a *dominance principle*[6]) that f1 is worse than f2 and that c1 is wrong. Let's call those claims together the standard view. If the standard view is true, then that would seem to suggest – and I admit that there's a leap in the argument at this point[7] – that f3 is worse than f4 in the following case (Simple Addition; Graph 7). Still, that leap of logic seems at least plausible since it seems that any plausible theory that implies that f1 is worse than f2 though worse for no one also will imply that f3 is worse than f4 though worse for no one.

Graph 7: Simple Addition

	c3	c4
probability	1	1
wellbeing	f3	f4
+10		**p**
+8		
+0	*p**	

Reply: We should reject the standard view, accept that f1 is exactly as good as f2 and that f3 is exactly as good as f4 – and that c1–c4 are all permissible.

Notably, this is *not* to say that we should also accept that the choice of depletion is permissible or that the future that includes depletion is exactly as good as the future that includes conservation. Ditto the "do nothing" choice in respect of climate change. The *simple* nonidentity problem is very different from the *probabilistic* form of the nonidentity problem, which includes depletion, risky policy, climate change, historical injustice, pleasure pill, slave child, etc. (See Part VII below.)

VI. Argument from additivity

We need additivity to insure correct logical features for the betterness relation (e.g., transitivity; no cycling) and also to make the evaluation process efficient (to insure the consistency of sequential pairwise comparisons). Additivity implies Pareto plus.

Reply: I suspect we can preserve the correct logical features – including transitivity and symmetry – without additivity. In any case, additivity doesn't imply Pareto plus. What we add up, under additivity, needn't be raw, unadjusted wellbeing levels. We can instead add up wellbeing adjusted to reflect our existential values.

6 Arrhenius (2000).
7 Person-affecting principles discussed by Bader and Holtug avoid this leap of logic. See Bader (unpublished ms.) and Holtug (2010).

Thus Broome argues that the *personal good* – what is added up to determine the overall (*general*) good of a world – can be understood to reflect the values of equality and fairness. Setting those values aside, I propose that personal good can be understood to reflect – that wellbeing can be adjusted to reflect – our existential values.

To make that proposal work in the Three Outcome Case (just to take one example), we would understand the personal good level for p in f3 where p's wellbeing is maximized to be zero and the personal good level for p in f2 p's wellbeing is avoidably lower to fall in the negative range (even though wellbeing itself is positive). In contrast, in the Two Outcome Case, the personal good level for p in f2 would just be zero.

VII. Argument from better (greater) chance

A better, that is, greater, *chance* of existence makes things morally better. At least: it's incontrovertible that Harry's better chance of existence under c1 makes c1 permissible. Therefore so must the *fact* of existence make things morally better.

Graph 8: Better Chance Case

	c1: take fertility pill		c2: take aspirin	
probability	0.1	0.9	0.0001	0.9999
wellbeing	f1	f2	f3	f4
+10			**Harry**	
+8	**Harry**			
+0		*Harry**		*Harry**

Reply: We should agree that Harry's better chance of existence under c1 makes the otherwise wrong c1 *permissible*.

That seems a core component of the person based intuition. But it's not one that EVNPBI itself expresses. The question is whether we can revise EVNPBI in a way that makes the point that the better chance at existence makes things better – at least in the sense that it makes the otherwise wrong c1 permissible – but avoids the further point, a point entirely at odds with the person based idea, that existence doesn't make things better. If we can't do that, we have an intuition at odds with itself.

Reply: We can do that. First a definition:

> Where a choice c made at a future x creates a probability n that p will have the wellbeing level (*wb*) that x in fact assigns to p, we can say that *x's probable value (PV) for p under c* is n(wb).
>
> (Note: We might naturally want to reference the fact that c1 creates more expected value for Harry than c2 does to explain c1's permissibility. But other cases show that approach fails by virtue of the fact that expected value can be greatly inflated by a low probability of a very wonderful outcome.[8])

And then:

> *Expansive Very Narrow Person Based Intuition + Probable Value (EVNPBI+PV)*: A future x is worse than a future y only if there is a p and an alternate accessible future z such that:
>
> p does or will exist in x *and*
> x is worse for p than z.
>
> c at x is wrong, only if there is a person p and an alternate choice c′ at an alternate accessible future y such that
>
> p does or will exist in x *and*
> x is worse for p than y *and*
> PV of c at x for p < PV of c′ at y for p.

To sum up: according to EVNPBI+PV, Harry's better chance under c1 at f1 makes c1 permissible (when it would otherwise be wrong). But, also according to EVNPBI+PV, c2 is permissible. Moreover, nothing in EVNPBI+PV suggests that Harry's better chance of existence embedded in f1 under c1 makes f1 better than f3 – makes, that is, make f3 worse than f1. So we are free to retain same-people Pareto and insist that f1 is worse than f3.

We are free, that is, to insist, more generally, that, where x and y assign to p the same wellbeing level but chances of p getting that wellbeing level – e.g. chances of p coming into existence at that level – are greater at x than at y, x is exactly as good as y.

Compare cases that ground the probabilistic (non-simple) versions of the non-identity problem, e.g., depletion, risky policy, "do nothing" choice on climate

8 Roberts (unpublished ms.). I owe this point to Dean Spears.

change, historical injustice, slave child, pleasure pill etc. In those cases it's clear we want to say the choice is wrong. And in those cases – on closer inspection – we should understand that the *apparent* victim – that is, the *victim* – doesn't have a better chance of existence under the suspect choice than he or she does under some (indeed many) relevant alternative choices.[9] Thus p's chances of existence are no greater under the choice of the pleasure pill than under the choice of the aspirin (the pleasure pill isn't a fertility pill); no greater under the choice of depletion, risky policy, than under the choice of conservation, safe policy.[10]

Compare the simple nonidentity problem (Graph 6 above), e.g., the case involving a genetic disorder that can't accessibly be mitigated or cured, cases in which the apparent victim can't be made better off (the agents lack the ability, the physical laws governing the world lack the flexibility), that is, in which the only alternative accessible future is one in which a nonidentical but better off child exists in place of the one.

VIII. Other issues

(1) Does EVNPBI+PV violate *complementarity* (Spears, unpublished ms.)? According to complementarity, it's better to assign (a "prospect" is better if it assigns) a higher probability of existence to the better off person than a lower probability of existence to the worse off person. By implication, it's better to assign a higher probability to the one person's existing and being better off than a lower probability to the one person's existing and being worse off. A case that would have those features is the fertility case amended so that the fertility pill is actually *safer* from the perspective of the future child than the aspirin is (perhaps now the aspirin causes brain bleeds).

Reply: No reason to think EVNPBI+PV violates complementarity. It *avoids* the implication that c2 at f3 is permissible. And we can still say that f1 is exactly as good as f2 is exactly as good as f4 and that all are better than f3.

---

9 That won't necessarily be the choice the agents *would have made* instead; the choice agents *would have made* in a given case may well make things worse for all involved. The choice we should compare against is instead one that agents could have made (whether they would have or not). Thus in the pleasure pill case we can add that, had the couple not taken the pleasure pill, they would, in a snit, not have conceived any child at all. That new fact is, however, a red herring – it doesn't change the analysis. The important thing is what the couple could have done. They could have taken the aspirin and conceived a child on just the same schedule on which they could have taken the pleasure pill and conceived a child. Note that in both cases, prior to choice, there exist a vast numbers of ways of implementing the particular choice, whether the choice to take the pleasure pill or the choice to take the aspirin. That's why under the pleasure pill choice as well any particular child's coming into existence is very low.

10 That is, the probable value the choice of the pleasure pill creates for the apparent victim isn't higher, but rather lower, than the choice of the aspirin creates for that same person.

(2) Does EVNPBI+PV violate *anonymity* (as described by Spears (unpublished ms.), or by Thomas (unpublished ms.), or by Bader (unpublished ms.), in the form of *impartiality*? Consider the following case:

Graph 9: Identity Matters

	c1	c2	c3
probability	1	1	1
wellbeing	f1	f2	f3
+10			**q**
+5	**p**	**q**	
+0	*q**	*p**	*p**

Anonymity implies that f1 is exactly as good as f2. But EVNPBI+PV implies that f1 is exactly as good as f3 and same-people Pareto that f3 is better than f2. Thus, f1 is better than f2, which violates anonymity.

Reply: We should accept the violation – that is, we should reject anonymity. Doing so preserves the highly intuitive EVNPBI+PV. Moreover, rejecting anonymity enables us to avoid implausible results in many infinite population problems. See e.g. Thomas's discussion of Welfarist Anonymity and Population Separability in the context of the case of "Steve." Thomas (unpublished ms.). The range of infinite population we shall ultimately need to address is considerable – and thus a discussion we shall need to reserve for another paper.

IX. Conclusion

In this paper, I have addressed a handful of interesting and often compelling arguments in favor of Pareto plus, the idea that other things equal the additional worth-having existence makes things morally better. I have argued that none of those arguments in fact succeeds in establishing that conclusion. The upshot is that it is beginning to look like we face no compelling reason not to reject Pareto plus if that is what we want to do. That result has practical importance. We retain credibility when we put forward analyses in support of the view that substantial sacrifice is both necessary and desirable for purposes of addressing climate change – that agents today ought to make, and be willing to impose on others, such sacrifice in order to make things go better for future generations. But we may lose credibility when we argue that agents ought not just to take steps to make things better for future generations but also take steps -- make, and impose on others, vast *additional* sacrifice – so that ever more future people can be brought into existence. If theory and credibility align – one allowing us to reject Pareto plus, the other simply a matter

of people's attitudes toward the sacrifices they may reasonably be asked to make – the specter of our utter failure to address climate change in the very near future may begin to recede.

References

Arrhenius, Gustaf (2000), "An Impossibility Theorem for Welfarist Axiologies," *Economics and Philosophy* 16: 247-266.

Arrhenius, Gustaf (2011), "The Impossibility of a Satisfactory Population Ethics," *World Scientific Review* 20(54): 1-26.

Bader, Ralf (forthcoming), "Person-Affecting Utilitarianism," in T. Campbell, K. Bykvist, G. Arrhenius, E. Finneron-Burns (eds.), *The Oxford Handbook of Population Ethics*. Oxford: Oxford University Press.

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

Roberts, Melinda A. (unpublished ms.), *The Existence Puzzles*.

Singer, Peter (2011), *Practical Ethics* (3rd. ed.), Cambridge University Press.

Spears, Dean (unpublished ms.), "Probabilistic Future People".

Thomas, Teruji (forthcoming), "Separability," in T. Campbell, K. Bykvist, G. Arrhenius, E. Finneron-Burns (eds.), *The Oxford Handbook of Population Ethics*. Oxford: Oxford University Press.

Anders Herlitz[1]

Nondeterminacy and Population Ethics

This paper synthesizes a general view out of Derek Parfit's last views on how to avoid the *Repugnant Conclusion* and presents the general features of a plausible theory of population ethics based on Parfit's suggestions. The paper argues that a plausible population axiology provides only partial orderings and implies that some outcomes are nondeterminate in their ranking. The paper shows, first, how the combination of what Parfit calls "imprecise equality" and the "Wide Dual Person-Affecting Principle" allows one to avoid both the *Continuum Argument* and the *Improved Mere Addition Parado*x. Second, the paper shows how this is enough to in principle also refute Gustaf Arrhenius's impossibility theorems. Third, the paper suggests that a plausible population axiology must allow for nondeterminacy, that whatever the substance of the axiology is, it can only provide partial orderings of outcomes, and that if we revise Arrhenius's adequacy conditions these can condition what a satisfactory population axiology looks like. Finally, the paper illustrates how one can apply normative theories that allow for nondeterminacy and also infer formal constraints on the theories in light of the consequences of their application.

[1] Institute for Futures Studies, anders.herlitz@iffs.se. Financial support by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences) is gratefully acknowledged.

Derek Parfit has done more than anyone to identify notorious problems in population ethics. One of the most pressing issues that he has drawn our attention to is the difficulty of avoiding the so-called *Repugnant Conclusion* (Parfit 1982, 1984):

> *The Repugnant Conclusion*: Compared with the existence of many people whose quality of life would be very high, there is some much larger number of people whose existence would be better, even though these people's lives would be barely worth living (Parfit 2016: 110, 2017: 153).

To grasp just how counterintuitive this position is it might help with some illustration. The Repugnant Conclusion, for instance, implies that compared with a population in which ten billion people flourish and lead lives filled with happiness, love, friendship and successful pursuit of worthwhile goals such as artistic creation, there is a much larger population in which all individuals lead lives which are barely worth living which is better. How could it be that a population – no matter its size – in which each individual has no other pleasure than "muzak and potatoes" (Parfit 1984) or the equivalent of the benefits accrued to a "lizard basking in the sun" (Parfit 2016) is better than a very large population in which everyone flourishes? To take this position seems absurd, and to suggest that a world filled only with people with damp lives is better than a world filled only with people with happy lives seems repugnant. The fact that the former population – with people leading damp lives – might be much, much bigger does not seem to affect our considered judgment that the latter population – with billions of people in happiness – is better.

Although the conclusion that the population with many damp lives is better to many indeed appears repugnant, several very compelling arguments support it. In particular, Parfit – and many with him – is troubled by two arguments that seem to give the view expressed in the Repugnant Conclusion very firm support: the *Continuum Argument* and different versions of the *Mere Addition Paradox*.

The Continuum Argument aims to establish that one must accept the Repugnant Conclusion if one makes pairwise comparisons of different populations the only difference between which is that one is significantly smaller but the individuals in it have *slightly* better lives. The argument relies on the following main premise:

> φ: Compared with the existence of many people who would all have lives that were equally worth living, there are some much larger numbers of people whose existence would be better, though these people would all have lives that would be slightly less worth living (Parfit 2016: 116).

This premise has significant intuitive support. To see this, compare the following populations. Population A consists of ten billion people who all have lives that are perfectly healthy and filled with happiness and meaningful activities. Population B consists of one hundred billion people who would all have lives that are filled with happiness and meaningful activities, but which are only almost perfectly healthy; the individuals in Population B all experience getting the flu one of the many winters they experience, and they all go through the quite unpleasant experiences associated with that. Intuitively, the fact that all the people in Population B need to endure a strike of the flu does not mean that Population A is better than Population B. Population B is ten times as big and the wellbeing of the people in Population B is nearly as good as the wellbeing of the people in Population A. Similar thought experiments seem possible for all population sizes and all levels of wellbeing, and thus the general claim expressed in φ seems very plausible.

If φ is true and if the better than relation is transitive (something some prominent philosophers have disputed, but which I will assume in this paper, cf. Rachels 1998; Temkin 1987, 2012), one must accept the Repugnant Conclusion. To see this, consider a continuum of populations, A-Z, in which the first population, A, consists of many people with very good lives, the second population, B, is much larger but the people in it have slightly worse lives, Population C consists of much more people than Population B, but the people in it have slightly worse lives, and so on until we reach Population Z which is an astronomically large population with lives that are barely worth living. Since φ tells us that an immediately successive population in this continuum is better than its predecessor, and since better than is a transitive relation, we must conclude that Z is better than A. Consider a simple illustration in which the width of the blocks shows the size of the populations and the height shows the quality of people's lives:

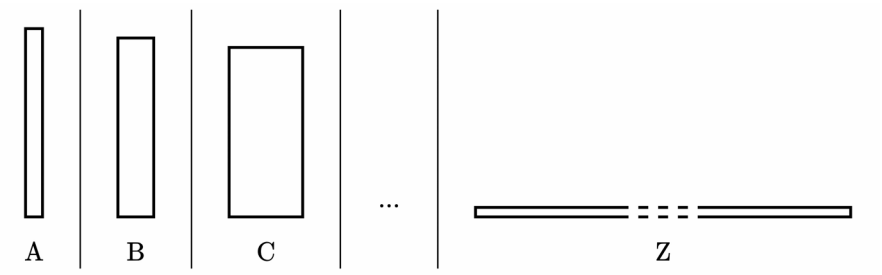


Figure 1

By only assuming that there is some population (however big) with individuals that lead lives that are only slightly worse that is better than a population with many people with good lives and that the better than relation is transitive, we are forced to conclude that A is worse than Z. That is the Repugnant Conclusion.

In his early writings on this topic, Parfit also expressed worries that the Repugnant Conclusion could not be avoided because of what he called the Mere Addition Paradox (Parfit 1982, 1984; Temkin 1987). I will here focus on a particular version of the Mere Addition Paradox. I take this version to be a stronger version of the initial argument, and it is the version Parfit discusses in his later writings (Parfit 2016). I will call this the *Improved Mere Addition Argument*. This argument resembles the Continuum Argument in that it makes use of the transitivity of the better than relation and comparisons of populations that are relatively close to each other in a continuum like the one described above. Contrary to the Continuum Argument, however, it does not rely merely on intuitive support for premise φ. Instead, it provides reasons to accept φ that are supported by other intuitions, expressed in premises π and ω:

> π: Compared with the existence of many people who would all have lives that were equally worth living, there is some larger population that consists of the very same people with even better lives as well as some additional people with lives that are slightly worse than the lives of the people in the first population that would be better.

Contrary to the main premise of the Continuum Argument, this premise does not appeal to intuitions about trade-offs between the number of people living and the quality of their lives. Instead, π only states that it would be better if the lives of everyone would be better even if this means that the population grows so that there is an additional number of people with lives that are slightly worse than the first group. This seems indisputable. Consider, for instance, Populations A and B above, where the only difference is that the people in Population B contracts the flu at some point in their lives. It seems ludicrous to deny that if one could improve the lives of all the people in Population A so that they would be even better by adding all the people in Population B and thereby create Population A+ then Population A+ would be better than Population A. Similar thought experiments seem possible for all population sizes and all levels of wellbeing, and thus the general claim expressed in π seems very reasonable.

> ω: Compared with the existence of many people with very good lives and many people with slightly worse lives, there is some population that consists of the same people where all lives are equally good, and where the total amount of wellbeing is greater that would be better.

This premise appeals to both inequality aversion and the preference for increasing the total amount of wellbeing. The underlying idea is that a larger amount of wellbeing that is more equally distributed is preferable to a smaller amount of wellbeing less equally distributed. Since the inequality aversion here is expressed in a very weak form (more equal distributions are favored when this increases total amount of wellbeing), this premise should be appealing to those who value equality, to those who value wellbeing increases and to those who value both equality and wellbeing increases. ω also appears intuitively plausible. Consider Population A+ again (the improved Population A together with Population B). Compared to A+, a population that consists of one group of people who lead phenomenally good lives and who never contracts the flu and a different group of people who lead phenomenally good lives but who contracts the flue and thus are sick for a couple of days, an alternative population that consists of both of these groups of people but where no one contracts the flu but everyone gets a minor stomach upset for a couple of hours at one point in their lives is better. Again, similar thought experiments seem possible for all population sizes and all levels of wellbeing, and thus the general claim expressed in ω seems very reasonable.

Accepting π and ω amounts to accepting φ. To see this, consider Figure 2:

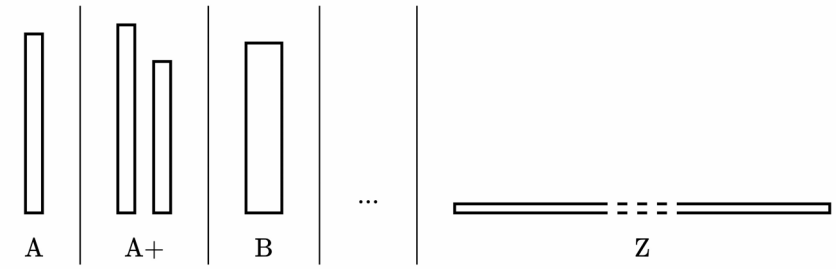


Figure 2

By accepting π and ω one accepts that there is some population with some number of lives (however big) with only slightly worse lives that is better than a population with many people with good lives and that the better than relation is transitive, we are forced to conclude that A is worse than Z. The difference in this argument is that

one does not move directly from A to B, from B to C, etc., but from A to A+ to B, from B to B+ to C, etc. Yet, the result is the same: A is considered worse than Z, which is the Repugnant Conclusion.

There are several proposals of how to avoid the problems that Parfit identifies in the large body of literature that comments on Parfit's arguments (for an overview, see: Arrhenius, Ryberg & Tännsjö 2010). Some have bitten the bullet and reject that the Repugnant Conclusion is at all repugnant (Tännsjö 2002). A supporting argument for that view might be that we cannot trust our intuitions in cases that involve large numbers (Broome 2004, but see also Pummer 2013 and Temkin 2012). Others have argued that there is some critical level at which an individual's wellbeing no longer contributes positively to the overall value of a population (Blackorby, Bossert & Donaldson 1995, 1997; Qizilbash 2007). By introducing such a critical level, one can reject the *general* validity of φ and ω and say that these premises are only valid for comparisons of certain populations, when individuals have wellbeing above the critical level. Yet others have suggested that the positive contribution of an individual's wellbeing to the overall value of a population is diminishing the more people exist in a population (Ng 1989). On this view, φ and π are rejected on the basis that both these premises falsely assume that the value of a person's wellbeing is independent of how many people exist, and at some point in the continuum the addition of more people will imply a smaller overall value of the population since it entails a smaller value for the wellbeing of a large number of people. Finally, some have suggested that the take-home lesson rather is that there can be no satisfactory population axiology (Arrhenius MS), i.e. no "betterness ordering of states of affairs, where the states of affairs include ones in which different numbers of persons are ever born" (Greaves 2017: 1).

Neither of these proposals is particularly appealing. Without going into too much detail, it can be noted that they all rely on accepting some deeply counter-intuitive view. Accepting that the Repugnant Conclusion is in fact not repugnant at all but just the result of sound inference of the application of appealing normative views goes against almost unanimous considered judgment. Introducing a critical level at which wellbeing no longer contributes positively to the value of a population implies a radical and implausible difference between wellbeing levels which are almost identical (cf. Arrhenius 2004). Depending on how one discounts the value of an individual's wellbeing, this view has very problematic implications. For example, Yew-Kwang Ng's proposal implies that for any number of tormented lives, there are situations in which it is better to add them rather than some people with very good lives (Arrhenius 2000). Rejecting the possibility of a satisfactory population axiology, finally, seems rash since we clearly seem able to order some states of affairs in which different number of people are ever born.

In the following sections, I will outline what I take to be a plausible way of avoiding the Repugnant Conclusion based on Parfit's latest publications on this issue. The second section addresses the possible objection that this proposal fails to deal with even more problematic issues in population ethics, in particular the so-called impossibility theorems. Finally, I discuss the practical implications of the Parfitian view and identify some ways of developing it further.

I

Building on two of Parfit's most recent papers, one can develop a general approach to how to avoid the Repugnant Conclusion. This approach has two fundamental elements. First, Parfit's argument that it is a mistake to accept what I will call the *Determinate Trichotomy Thesis* when thinking about the goodness of different outcomes (Parfit 2016). Second, Parfit's proposed substantive principle of what can make an outcome better than another: the *Wide Dual Person-Affecting Principle* (Parfit 2017).

Consider, first:

> *The Determinate Trichotomy Thesis*: If *x* and *y* are comparable with respect to *p*, it is determinately true that *x* is more *p* than y, that *x* is less *p* than *y*, or that *x* and *y* are equally as *p*.

The veracity of this thesis is often taken for granted. It is also often tacitly assumed when the premises used in the Continuum Argument and in the Improved Mere Addition Paradox are considered. When comparing Population A with Population B, it is often thought that if one rejects that B is determinately better than A, then one must be committed to thinking that in so far as A and B at all can be compared with respect to how good they are, B is determinately worse than A, or A and B are determinately equally good. This would follow if the Determinate Trichotomy Thesis were true.

However, the Determinate Trichotomy Thesis has been increasingly questioned lately. Some have argued that the *trichotomy* better than, worse than and equally as good as does not exhaust the set of possible positive value relations (Carlson 2010; Chang 2002, 2016). On this view, it might be true that *x* and *y* are comparable with respect to *p*, but not true that either is more *p* than the other or that they are equally as *p*. Instead, proponents of this view hold that some non-conventional comparative relation such as *parity* might obtain between *x* and *y* with respect to *p*. Others question the Determinate Trichotomy Thesis because they believe that it is some-

times *indeterminate* whether *x* is more *p* than *y*, less *p* than *y* or equally as *p* as *y* (Broome 1997, 2004; Contantinescu 2014; Dougherty 2014; Elson 2017). Indeterminacy of this kind is possible if *p* is vague.

Parfit rejects the Determinate Trichotomy Thesis but presents a different kind of argument for this. He argues that it is sometimes the case that *x* and *y* are *imprecisely* equally as *p* (which is to be distinguished from being equally as *p*). What Parfit means with this is somewhat unclear, but it is clear that he accepts that there are only three kinds of positive relations and he does not think that imprecision can be explained with vagueness:

> [T]here are only imprecise truths about the relative goodness of many different acts or outcomes, such as ones that would greatly benefit a few people, or give lesser benefits to many others. Such imprecision is not the result of vagueness in our concepts, or our lack of knowledge, but is part of what we would know if we knew the full facts. When two things are qualitatively very different, these differences would often make it impossible either that one of these things is better than the other by some precise amount, or that both things are precisely equally good (Parfit 2016: 113).

What is important is that contrary to the relation equally as *p*, imprecisely equally as *p* is not transitive. If *x* is equally as *p* as *y*, and *y* is equally is *p* as *z*, then *x* and *z* are equally as *p*. If *x* is imprecisely equally as *p* as *y*, and *y* is imprecisely equally is *p* as *z*, then *x* is not necessarily imprecisely equally as *p* as *z*.

The introduction of non-transitivity is something Parfit's proposal shares with the non-conventional comparative relations and the indeterminacy proposals, and as we will see this is also a key to avoiding the Repugnant Conclusion. Parity is by definition non-transitive (Chang 2002). Those who believe that it is sometimes indeterminate whether *x* is more *p* than *y* rejects transitivity in a more subtle way. They do not defend non-transitive comparative relations, but they believe that the following is possible: If it is indeterminate whether *x* is equally as *p* as *y*, and indeterminate whether *y* is equally is *p* as *z*, then it is not necessarily the case that it is indeterminate whether *x* is equally as *p* as *z*.

Rather than espousing Parfit's particular view and accepting the possibility of imprecise equality, I believe that it is more fruitful to focus on the negative view, the rejection of the Determinate Trichotomy Thesis. There are competing explanations for *why* the Determinate Trichotomy Thesis must be rejected (e.g. the possibility of parity, indeterminacy, impreciseness), but how one explains this has little impact on what a plausible population axiology looks like. The take-home lesson is that rejecting the Determinate Trichotomy Thesis might help one solve problems in

population ethics. I will in what follows say that someone who rejects the Determinate Trichotomy Thesis accepts *nondeterminacy*, and accepts that alternatives sometimes are *nondeterminate in their ranking* (cf. Herlitz 2019). Everyone who rejects the Determinate Trichotomy Thesis should accept nondeterminacy, regardless of why they reject the Determinate Trichotomy Thesis.

Rejecting the Determinate Trichotomy Thesis has significant implications. Consider what we, following Joseph Raz's and Muzaffar Qizilbash's discussions of incommensurability and parity and the widespread use of the so-called small improvement argument in the literature on parity, might call a *Mark of Nondeterminacy* (cf. Chang 2002; Qizilbash 2007; Raz 1986):

> *A Mark of Nondeterminacy*: If x and y are nondeterminate in their ranking with respect to p, then an improvement (worsening) in x will not necessarily make the resulting x more (less) p than y.

A characteristic feature of nondeterminacy is that if it is true that x is not determinately worse than y, and also true that y is not determinately worse than y, then an improvement in x does not necessarily make the resulting x ($x+$) better than y. This distinguishes nondeterminacy from equality. If x and y are determinately equally as good, then an improvement in x will necessarily make the resulting x ($x+$) better than y. This is so because equally as good as is a transitive relation so if $x+$ was equally as good as y it would be equally as good as x, but it is by definition better than x.

Thus, by rejecting the Determinate Trichotomy Thesis Parfit rejects the idea that transitive relations that apply determinately can always rank alternatives with respect to how good they are. This entails accepting the implications of the Mark of Nondeterminacy, that sometimes when neither of two alternatives is determinately worse than the other and one introduces an improvement in one of them, this does not necessarily result in that the resulting, improved alternative is better than the other.

Consider, second:

> *The Wide Dual Person-Affecting Principle*: One of two outcomes would be in one way better if this outcome would together benefit people more, and in another way better if this outcome would benefit each person more (Parfit 2017: 154).

Some explanation of this principle is required. The context in which Parfit presents this view is a discussion of the so-called *Non-Identity Problem*. Like the Repugnant

Conclusion, this is a notorious problem in population ethics that was identified by Parfit in his earlier work and which has generated a large body of literature (Parfit 1984; Roberts 2015). A fundamental question in the debate on the Non-Identity Problem is whether something can be good without being good *for* someone, and in particular whether some benefit can be good if the person benefitting would not exist had it not been for the benefit. In his last paper on this issue, Parfit espouses a view presented by Jeff McMahan:

> If someone is caused to exist and to have a life that is worth living, that is good for this person, giving him or her an existential benefit. There are similar existential harms (McMahan 2013: 6-7).

This is what Parfit has in mind when he claims that his principle is a *wide person-affecting* principle (as opposed to a *narrow* person-affecting principle which rejects the possibility of existential benefits). It is a principle according to which nothing can be good (bad) unless it is good (bad) *for* someone, but this is understood so that it can be good (bad) for a person to be caused to exist.

Next, Parfit distinguishes between two senses in which one of two outcomes might benefit people more. An outcome would:

> (1): benefit people more in the *collective* sense if this outcome would *together benefit people more*.
>
> (2): benefit people more in the *individual* sense if this outcome would *benefit each person more* (Parfit 2017: 152).

Parfit believes that an outcome or population can be better than another in both of these senses. First, he accepts that if one of two outcomes contains a greater total sum of wellbeing, it is in one way better; it benefits people more in the collective sense. Second, he accepts that if one of two outcomes benefits each person more and thereby is better for every individual in the outcome, it is in one way better; it benefits people more in the individual sense. Thus, the Wide Dual Person-Affecting Principle states that one of two outcomes would be in one way better if the total amount of wellbeing is greater, and in another way better if this outcome would benefit each person more.

It is clear that some outcomes are better than others in the collective sense but not in the individual sense. This is true, for instance, of the populations in the continuum of populations that are compared in the Continuum Argument. Population

B is better than Population A in the collective sense, but not in the individual sense.

Importantly, if one – like Parfit – rejects the Determinate Trichotomy Thesis one is not committed to the idea that we can always determine which of two outcomes is at least as good as the other *overall* when one outcome is better than another in one respect but not in the other respect. Instead, one can accept that two outcomes might be nondeterminate in their ranking with respect to the collective and the individual senses of benefit. In other words, the Wide Dual Person-Affecting Principle allows for nondeterminacy and fails to always fully determine an outcome that is at least as good as all alternative outcomes.

In combination, rejecting the Determinate Trichotomy Thesis and accepting the Wide Dual Person-Affecting principle tell us that: an outcome is (i) in one way better than another if it would together benefit people more and (ii) in another way better if it would benefit each person more, but (iii) we should not assume that we can always fully determine that any of the following relations obtain between two outcomes: better than, worse than, equally as good as.

We now have the tools to dismiss both the Continuum Argument and the Improved Mere Addition Paradox, and what transpires is a thus a view that seems to avoid the Repugnant Conclusion. Both the Continuum Argument and the Improved Mere Addition Paradox rely on the tacit assumption that one can rank the populations under consideration with transitive relations, so rejecting this is in itself sufficient to dismiss both these arguments. However, in order to explain why the premises that these arguments rely on are appealing and how this can be accounted for, one needs a substantive view of what makes an outcome better than another. The Wide Dual Person-Affecting Principle is such a view.

Consider first the Continuum Argument. This argument can be dismissed on the grounds that it mistakenly assumes that if an outcome is significantly better than another in one way, it can always be determined that it is all things considered better. For some populations next to each other in the continuum, we might say that it will be false that either is at least as good as the other. Instead, it will be the case that they are nondeterminate in their ranking. These populations next to each other in the continuum will be nondeterminate in their ranking because one will be better than the other in the collective sense, but not in the individual sense, and in some situations it is impossible to fully determine that either of two outcomes is at least as good as the other with respect to these considerations. Since nondeterminacy is a non-transitive relation, this breaks the series of inferences that leads one from holding that B is better than A to Z is better than A. It allows one to, for instance, hold the view that A is worse than B, that B and C are nondeterminate in their ranking, but that A is better than C.

Similarly, the Improved Mere Addition Paradox can be dismissed with reference

to how one of its premises (ω) mistakenly assumes that if an outcome is significantly better than another in one way, it can always be determined that it is all things considered better. For some populations with features that are similar to Populations A+ and B, where the populations are equally big, the first population contains inequality and the second population is perfectly equal, contains more wellbeing, but the best off are worse off, it will be false that either is at least as good as the other. Instead, it will be the case that they are nondeterminate in their ranking. These populations next to each other in the continuum will be non-determinate in their ranking because one will be better than the other in the collective sense, but not in the individual sense, and in some situations it is impossible to fully determine that either of two outcomes is at least as good as the other with respect to these considerations. Again, since nondeterminacy is a non-transitive relation, this breaks the series of inferences that leads one from holding that B is better than A to Z is better than A.

The issue of where on the continuum the nondeterminacy occurs relates to how one specifies the Wide Dual Person-Affecting Principle. Parfit provides no specification of this principle, but it can be recognized that there is a wide range of possible specifications compatible with Parfit's general proposal. There are at least three dimensions in which the Wide Dual Person-Affecting Principle can be specified: (i) what is the relative value of collective benefits compared to individual benefits; (ii) how wide is the scope of nondeterminacy, i.e. how many outcomes will the principle fail to determinately rank trichotomously; (iii) where on the continuum does the nondeterminacy occur, i.e. does it arise when collective benefits are high (low) and/or when individual benefits are high/low?

A general conclusion can now be drawn by a Parfitian approach to how one can avoid the Repugnant Conclusion: A population axiology can avoid the Repugnant Conclusion while it accounts for the underlying intuitions that otherwise lead to the Repugnant Conclusion if these intuitions are reflected in a principle that allows for nondeterminacy. This is compatible with various views of why nondeterminacy arises. This might be due to non-conventional comparative relations such as parity, vagueness, or what Parfit referred to as impreciseness. It is also compatible with a variety of substantive principle, including various specifications of the Wide Dual Person-Affecting Principle.

II

In a paper that comments on Parfit's proposal of how to avoid the Continuum Argument and the Improved Mere Addition Paradox by rejecting the Determinate Trichotomy Thesis, Gustaf Arrhenius raises another concern: even if one rejects the

Determinate Trichotomy Thesis, it is not clear how one can deal with the impossibility theorems that show that there can be no satisfactory theory of population ethics (Arrhenius 2016). In this section, I outline a response to Arrhenius based on the Parfitian approach to how to avoid the Repugnant Conclusion.

Arrhenius has presented six different theorems in which he shows that no population axiology can meet a set of highly plausible adequacy conditions (Arrhenius MS). Of these, Arrhenius believes that the sixth theorem is the strongest (Arrhenius 2016). I will here focus on the strongest theorem:

> *Arrhenius's Sixth Impossibility Theorem*: There is no population axiology which satisfies the Egalitarian Dominance, the General Non-Extreme Priority, the Non-Elitism, the Weak Non-Sadism, and the Weak Quality Addition Condition (Arrhenius 2009).

Arrhenius argues that there is no population axiology that satisfies the following adequacy conditions (here presented in the informal way):

> *The Egalitarian Dominance Condition*: If population A is a perfectly equal population of the same size as population B, and every person in A has higher welfare than every person in B, then A is better than B, other things being equal.

> *The General Non-Extreme Priority Condition*: There is a number *n* of lives such that for any population X, and any welfare level **A**, a population consisting of the X-lives, *n* lives with very high welfare, and one life with welfare **A**, is at least as good as a population consisting of the X-lives, *n* lives with very low positive welfare, and one life with welfare slightly above **A**, other things being equal.

> *The Non-Elitism Condition*: For any triplet of welfare levels, **A**, **B**, and **C**, **A** slightly higher than **B**, and **B** higher than **C**, and for any one-life population A with welfare **A**, there is a population C with welfare **C**, and a population B of the same size as A∪C and with welfare **B**, such that for any population X consisting of lives with welfare ranging from **C** to **A**, B∪X is at least as good as A∪C∪X, other things being equal.

> *The Weak Non-Sadism Condition*: There is a negative welfare level and a number of lives at this level such that an addition of any number of people with positive welfare is at least as good as an addition of the lives with negative welfare, other things being equal.

> *The Weak Quality Addition Condition*: For any population X, there is a perfectly equal population with very high positive welfare, and a very negative welfare level, and a number of lives at this level, such that the addition of the high welfare population to X is at least as good as the addition of any population consisting of the lives with negative welfare and any number of lives with very low positive welfare to X, other things being equal (Arrhenius 2009: 28-30).

Arrhenius proves that no population axiology can meet all five adequacy conditions. This holds also for the Parfitian view, and so one wonders which of the adequacy conditions to reject.

In light of the Wide Dual Person-Affecting Principle, three of the conditions seem indisputably valid. The Egalitarian Dominance condition states that if two populations have the same size, there is perfect equality in both populations, and the wellbeing of the people in one population is greater than the wellbeing of the people in the other, then the population with the better off people is better. The Wide Dual Person-Affecting Principle is in complete agreement with this. Since both the total amount of wellbeing is higher and everyone is better off in one of the populations, this population is better in both of the respects that the Wide Dual Person-Affecting Principle recognizes. Also, the Weak Non-Sadism Condition is clearly met by the Wide Dual Person-Affecting Principle, since this principle implies that adding people with negative wellbeing is bad and that adding people with positive wellbeing is good. Finally, the Weak Quality Addition Condition is met by the Parfitian view since it is a weaker formulation of avoidance of the Repugnant Conclusion (cf. Arrhenius 2016; Fehige 1998). Thus, even if the Wide Dual Person-Affecting Principle does not in itself necessarily meet this condition, one must be committed to a specification of this principle that meets the condition in so far as one wants to avoid the Repugnant Conclusion.

Those accepting the Parfitian view might, however, reject either the General Non-Extreme Priority Condition or the Non-Elitism Condition, or both. To see this, consider how both of these conditions restrain what a population axiology must say about comparisons of different distributions where it is the case that some are better off in one distribution and some are better off in a different distribution. The Parfitian view of the implications of the Wide Dual Person-Affecting Principle clearly is that in some cases in which some are better off in one outcome and some are better off in a different outcome the two outcomes are nondeterminate in their ranking. This is how one dismisses the Improved Mere Addition Paradox. Those accepting the Parfitian view might, thus, argue that either one or both of these conditions mistakenly requires of a population axiology that it implies that if an outcome is significantly better than another in some but not all ways, it can always

be determined that it is all things considered better. If one is inclined to believe that nondeterminacy arises when the worst off in a population is worse off in one of the outcomes, one might reject the General Non-Extreme Priority Condition. If one is inclined to believe that nondeterminacy arises when the best off in a population is worse off in one of the outcomes, one might reject the Non-Elitism Condition. If one believes that nondeterminacy is plausible in both kinds of situation, one might reject both conditions. The Wide Dual Person-Affecting Principle is compatible with all of these positions. Which one to go for depends on how one thinks the principle should be specified.

In light of this, one might suggest that Arrhenius makes a mistake when he formulates his adequacy condition by ruling out nondeterminacy. To allow for nondeterminacy, either or both of the following can instead be used as adequacy conditions:

> *The General Non-Extreme Priority Condition**: There is a number *n* of lives such that for any population X, and any welfare level **A**, a population consisting of the X-lives, *n* lives with very high welfare, and one life with welfare **A**, is *not determinately worse than* a population consisting of the X-lives, *n* lives with very low positive welfare, and one life with welfare slightly above **A**, other things being equal.
>
> *The Non-Elitism Condition**: For any triplet of welfare levels, **A**, **B**, and **C**, **A** slightly higher than **B**, and **B** higher than **C**, and for any one-life population A with welfare **A**, there is a population C with welfare **C**, and a population B of the same size as AUC and with welfare **B**, such that for any population X consisting of lives with welfare ranging from **C** to **A**, BUX is *not determinately worse than* AUCUX, other things being equal.

These definitions capture the general intuitions behind the adequacy conditions, aversion to extreme priority to the worst off and to elitism respectively, while they allow for nondeterminacy. However, they do not (together with the other adequacy conditions) rule out the possibility of a population axiology. To see this, it suffices to note that the Wide Dual Person-Affecting Principle can be understood so that it meets all five adequacy conditions if one replaces either the non-elitism condition or the non-extreme priority condition as suggested above. It can be noted that this revision is also in line with some other work on population ethics, in particular the approaches that want to allow for indeterminacy due to vagueness (Broome 2004; Qizilbash 2007, 2018; Thomas 2015).

It could of course be objected that these revisions of the adequacy conditions

imply that they lose significant intuitive support. The aversion to extreme priority to the worst off and to elitism indicates that there are some outcomes that are *better than* the outcomes in which the worst off/best off fares best. The revised adequacy conditions fail to reflect this.

In response to this objection, and in the spirit of Parfit's latest work, it is worth pointing out that the question we must ask is not whether it is in itself plausible to revise these conditions. Instead, the relevant question is whether it is *more implausible* to make these revisions than to accept that there can be no satisfactory population axiology. Making these revisions seems to me to be the least implausible of these options.

Furthermore, there might be further reasons to reject the Determinate Trichotomy Thesis and accept revisions of Arrhenius adequacy conditions of this kind. For the Determinate Trichotomy Thesis to be true it is not enough that the normative principles that apply are able to in principle provide determinate trichotomous orderings. It must also be possible to order the elements that these principles order with respect to their pertinent characteristics. In population ethics, two characteristics are generally considered pertinent when one considers different populations: size and wellbeing. Of course, size is measured by natural numbers and natural numbers by their nature allow for determinate, trichotomous orderings. It is determinately true for all pairs of natural numbers that one is greater than the other or that they are of equal size. It is, however, far from obvious that one can say the same about wellbeing. Instead, it might be the case that two lives are nondeterminate in their ranking with respect to wellbeing. This might, for instance, be a plausible conclusion when one compares one life that is short but filled with achievements and happiness with a life that is long but filled with health problems. If lives with different properties cannot be determinately, trichotomously ordered with respect to wellbeing, this challenges the assumption that populations can be determinately, trichotomously ordered with respect to how good they are. If one rejects the Determinate Trichotomy Thesis on the grounds that wellbeing levels cannot be determinately, trichotomously ranked, this gives one further reason to allow for nondeterminacy and revise Arrhenius's adequacy conditions.

III

The following general hypothesis can now be introduced in light of Parfit's approach to how to avoid the Repugnant Conclusion and the discussion of Arrhenius's impossibility theorem:

> *Nondeterminate Population Ethics*: An approach to population axiology can only be satisfactory if it allows for nondeterminacy.

It might be argued that one concedes too much by giving up of the Determinate Trichotomy Thesis since this implies that one loses the ability to make practical judgments with the normative principles one accepts. If one's principles only manage to establish that some alternatives are nondeterminate in their ranking, it is hard to see what this theory prescribes. There is a sense in which a theory that allows for nondeterminacy is unsatisfactory.

This worry is misguided. Depending on how one explains nondeterminacy, one can use principles that allow for nondeterminacy to weed out impermissible alternatives (cf. Herlitz 2019). Conceptualizing nondeterminacy in terms of incompleteness allows one to use Amartya Sen's conception of maximization and say that all alternatives that are not maximal (i.e. all alternatives that are determinately worse than some alternative) are impermissible (Sen 1997, 2017). Thinking of nondeterminacy in terms of vagueness enables one to say that all alternatives that are worse than some alternative on some admissible precisification are impermissible (Andersson 2017; Broom 2009; Fine 1975). Using the fitting attitudes approach to value, one might stipulate that all alternatives that are impermissible to prefer or equiprefer to some alternative are impermissible (cf. Rabinowicz 2008).

I will not suggest that any of these approaches is preferable to the others, but it is important to note that there are several ways in which principles that allow for nondeterminacy can be used to create partial orderings. This reveals a significant advantage that this approach has to skepticism about the possibility of a satisfactory population axiology. The skeptic cannot give any explanation of why a population with many people who lead mediocre lives is worse than a population with the exact same people in it with better lives. A principle that allows for nondeterminacy has no problems determining that it is better if everyone lead better life.

Nevertheless, a problem will remain for proponents of principles that allow for nondeterminacy: in some situations, two or more alternatives in the feasible choice set will remain after one has discarded impermissible alternatives. How to choose between alternatives that are nondeterminate in their ranking is no small problem. Should we consider them equally permissible on the ground that they all passed the impermissibility test? Or should we introduce some further condition? Following Sen, one might come to the conclusion that one can rationally choose any option that is not determinately worse than any alternative (Sen 1997). Following Ruth Chang, one might instead recognize that when two alternatives are nondeterminate in their ranking there is sometimes what Chang calls a "resolutional remainder" that implies that further reasons are needed to justify a choice (Chang 2002). When two

alternatives are nondeterminate in their ranking people can reasonably disagree about which is better, and reasons apply to this disagreement. This question must be given more attention by population ethicists who accept nondeterminacy.

Regardless of what one thinks about whether or not further reasons are needed to choose between alternatives that are nondeterminate in their ranking, there might be formal constraints that apply to such choices. To see this, consider a well-known problem that arises as soon as one rejects the Determinate Trichotomy Thesis: sequences of choice situations in which some alternatives are nondeterminate in their ranking can sometimes be determinately, trichotomously ranked even if alternatives in each step of the sequences cannot be determinately, trichotomously ranked (cf. Elga 2010; McClennen 1990).

Consider an illustration. Assume that A and B are nondeterminate in their ranking, that A is worse than A+, and that A+ and B are nondeterminate in their ranking. Assume separability so that the values of A, A+ and B in the two choice situations are independent of other choices:

	Choice 1	Choice 2
Sequence X	A	B
Sequence Y	B	A+

Table 1

We can easily see that Sequence X is determinately worse than Sequence Y since A+ is determinately better than A, and the only difference in value between the sequences is that one includes A and the other A+ (remember that we have assumed separability).

In light of this phenomenon, it can be suggested that whatever approach one takes to alternatives that are nondeterminate in their ranking, this approach should rule out the possibility of making sequences of choices that are determinately worse than some other sequence. However one formulates such a criterion, it constrains how to make decisions in face of nondeterminacy. For instance, it rules out the view that one can permissibly flip a coin between alternatives that are nondeterminate in their ranking, but it does not rule out the view that one should always favor A-type values over B-type values when some alternatives are nondeterminate in their ranking. Population ethicists that accept nondeterminacy should give more attention to what kinds of formal criteria one can pose of decision methods when the normative grounds for the decisions allow for nondeterminacy.

IV

Parfit's last work on population ethics teaches us that we can avoid the Repugnant Conclusion if we reject the Determinate Trichotomy Thesis and accept nondeterminacy. This also provides us with the tools necessary to pose relevant adequacy conditions on what a satisfactory population axiology is that allow for the possibility of a population axiology. A plausible population axiology is a population axiology that only partially orders outcomes. A particular question that arises when we accept this is: to what extent can one introduce formal constraints on what practical judgments are permissible with respect to a population axiology that allows for nondeterminacy? Nondeterminacy does not pose a general challenge to the possibility of forming practical judgments based on one's population axiology, but it means that we should think of practical judgments in population ethics in unconventional ways.

References

Andersson, Henrik. 2017. *How it all relates: Exploring the space of value comparisons*. Ph.D. dissertation, Lund University.

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

Arrhenius, Gustaf. 2004. "Superiority in Value," *Philosophical Studies*, 123: 97–114.

Arrhenius, Gustaf. 2009. "One More Axiological Impossibility Theorem," in Lars-Göran Johansson, Jan Österberg and Rysiek Sliwinski (eds.), *Logic, Ethics, and All That Jazz: Essays in Honour of Jordan Howard Sobel*. Uppsala: Uppsala University: 23–37.

Arrhenius, Gustaf. 2016. "Population Ethics and Different-Number-Based Imprecision," *Theoria*, 82(2): 166-181.

Arrhenius, Gustaf. MS. *Population Ethics: The Challenge of Future Generations*.

Arrhenius, Gustaf, Jesper Ryberg and Torbjörn Tännsjö. 2010. "The Repugnant Conclusion," in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*.

Blackorby, Charles, Walter Bossert and David Donaldson. 1995. "Intertemporal Population Ethics: Critical-Level Utilitarian Principles," *Econometrica*, 63(6): 1303-1320.

Blackorby, Charles, Walter Bossert and David Donaldson. 1997. "Critical-Level Utilitarianism and the Population-Ethics Dilemma," *Economics and Philosophy*, 13(2): 197–230.

Broome, John. 1997. "Is Incommensurability Vagueness?" in Ruth Chang (ed.), *Incommensurability, Incomparability, and Practical Reason*. Cambridge, MA: Harvard University Press.

Broome, John. 2004. *Weighing Lives*. Oxford: Oxford University Press.

Broome, John. 2009. "Reply to Rabinowicz," *Philosophical Issues* 19: 41–-417.

Carlson, Erik. 2010. "Parity Demystified." *Theoria*. 76, 119–128.

Chang, Ruth. 2002. "The Possibility of Parity," *Ethics*, 112(4): 659–688.

Chang, Ruth. 2016. "Parity, Imprecise Comparability and the Repugnant Conclusion," *Theoria*, 82(2): 182-214.

Contantinescu, Cristian. 2014. "Moral Vagueness: A Dilemma for Non-Naturalism," in Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics, Volume 9*. Oxford: Oxford University Press: 152–185.

Dougherty, Tom. 2014. "Vague Value," *Philosophy and Phenomenological Research*, 89(2): 352–372.

Elga, Adam. 2010. "Subjective Probabilities Should be Sharp," *Philosopher's Imprint* 10: 1–11.

Elson, Luke. 2017. "Incommensurability as Vagueness: A Burden-Shifting Argument." *Theoria*. 83: 341–363.

Fehige, Christoph. 1998. "A Pareto Principle for Possible People," in Christop Fehige and Ulla Wessels (eds.), *Preferences, Perspektiven der analytischen Philosophie; Perspectives in analytical philosophy*. Berlin: W. de Gruyter: 508–543.

Fine, Kit. 1975. "Vagueness, Truth and Logic," *Synthese*, 30: 265-300.

Greaves, Hilary. 2017. "Population Axiology," *Philosophy Compass*, 12(11): e12442.

Herlitz, Anders. 2019. "Nondeterminacy, two-step models and justified choice." *Ethics*, 129(2): 284–308.

McClennen, Edward F. 1990. *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge: Cambridge University Press.

McMahan, Jeff. 2013. "Causing People to Exist and Saving People's Lives," *Journal of Ethics*, 17(1): 5-35.

Ng, Yew-Kwang. 1989. "What Should We Do about Future Generations? The Impossibility of Parfit's Theory X," *Economics and Philosophy*, 5(2): 235-253.

Parfit, Derek. 1982. "Future Generations: Further Problems," *Philosophy and Public Affairs*, 11: 113–172.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Parfit, Derek. 2016. "Can We Avoid the Repugnant Conclusion?" *Theoria*, 82(2): 110–127.

Parfit, Derek. 2017. "Future People, the Non-Identity Problem, and Person-Affecting Principles," *Philosophy and Public Affairs*, 45(2): 118–157.

Pummer, Theron. 2013. "Intuitions about large number cases," *Analysis*, 73: 37–46.

Qizilbash, Mozaffar. 2007. "The Mere Addition Paradox, Parity and Vagueness," *Philosophy and Phenomenological Research*, 75(1): 129–151.

Qizilbash, Mozaffar. 2018. "On Parity and the Intuition of Neutrality," *Economics and Philosophy*, 34: 87–108.

Rabinowicz, Wlodek. 2008. "Value Relations," *Theoria*, 74: 18–49.

Rachels, Stuart. 1998. "Counterexamples to the Transitivity of Better Than," *Australasian Journal of Philosophy*, 76: 71–83.

Raz, Joseph. 1986. *The Morality of Freedom*. Oxford: Oxford University Press.

Roberts, M. A. 2015. "The Nonidentity Problem," in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*.

Sen, Amartya. 1997. "Maximization and the Act of Choice," *Econometrica*, 65: 745–779.

Sen, Amartya. 2017. "Reason and Justice: The Optimal and the Maximal," *Philosophy*, 92: 5–19.

Temkin, Larry. 1987. "Intransitivity and the Mere Addition Paradox," *Philosophy and Public Affairs*, 16: 138–187.

Temkin, Larry. 2012. *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. Oxford: Oxford University Press.

Thomas, Teruji. 2017. "Some Possibilities in Population Axiology," to appear in *Mind*. DOI:10.1093/mind/fzx047

Tännsjö, Torbjörn. 2002. "Why We Ought to Accept the Repugnant Conclusion," *Utilitas*, 14(3): 339–359.

Wlodek Rabinowicz[1]

Can Parfit's Appeal to Incommensurabilities Block the Continuum Argument for the Repugnant Conclusion?

Blocking the Continuum Argument for the Repugnant Conclusion by an appeal to incommensurabilities in value, as suggested in Parfit (2016), is an attractive option. But incommensurabilities ('imprecise equalities' in Parfit's terminology) that need to be posited to achieve this result have to be very thoroughgoing – 'persistent' in the sense to be explained. While this persistency is highly atypical, it can be explained if incommensurability is interpreted on the lines of the fitting-attitudes analysis of value, as permissibility of divergent attitudes towards the items that are being compared. More precisely, it can be interpreted as *parity* – as the permissibility of opposing preferences with respect to the compared items. This account makes room for the persistency phenomena. Nevertheless, some of Parfit's substantive value assumptions must be given up, to avoid implausible implications. In particular, his Simple View regarding the marginal value of added lives cannot be retained.

[1] Lund University, wlodek.rabinowicz@fil.lu.se. Financial support by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences) is gratefully acknowledged.

1. Introduction

Starting with *Reasons and Persons* (1984), Parfit repeatedly grappled with the Repugnant Conclusion. His last published attempt, "Can We Avoid the Repugnant Conclusion?," was published as late as in 2016, in a special issue of the Swedish international philosophy journal *Theoria*. The issue collected contributions from a symposium held in Stockholm in 2014 on the occasion of Parfit being awarded the prestigious Rolf Schock Prize in Logic and Philosophy. Having chaired the prize symposium I have long been drawn to the idea of commenting on Parfit´s proposal. This Festschrift in his memory provides a welcome opportunity to do so.

In his *Theoria*-paper, Parfit attempts to block the "Continuum Argument" for the Repugnant Conclusion by an appeal to "imprecise equalities" in value.[2] My aim is to assess this attempt. By imprecise equality Parfit means something very close to *incommensurability* – a relation that holds between two items if (and only if) none of them is better than the other nor are they (precisely) equally as good.[3] Or, at least, imprecise equality is meant to entail incommensurability.

> Precisely equal is a transitive relation. [...] But if X and Y are imprecisely equally good, so that neither is worse than the other, these imprecise relations are not transitive. [...] Two things are imprecisely equally good if it is true that, though neither thing is better than the other, there could be some third thing which was better or worse than one of these things, though not better or worse than the other. (Parfit 2016, pp. 14f)

Is the comparison with "some third thing" intended to be a part of the definition of imprecise equality? Or is it just meant to be a useful (partial) test that allows us to identify this relation, as it implies that X and Y aren't precisely equal in value? Here, I will assume the latter. Indeed, I will assume that imprecise equality is the same as

[2] The 'continuum' label, which is Parfit´s own, is not really adequate. The argument doesn't involve a continuum but only a discrete sequence of larger and larger populations, with gradually decreasing wellbeing level.

[3] To prevent misunderstandings, the reader should be warned that this definition of incommensurability, while nowadays quite common among value theorists (probably due to the influence of Raz 1986), is very different from how this notion has been traditionally understood. On the traditional understanding, two items are commensurable in value if their respective values can be measured on a common cardinal scale. They are incommensurable otherwise. Now, obviously, cardinal measurability of value is a strong requirement. It is therefore not surprising that Chang (2016), who interprets incommensurability in this traditional way, denies that Parfit´s imprecise inequality is the same thing as incommensurability. Two items might well be incommensurable in this traditional sense without being imprecisely equal. On my definition, only a common *ordinal* scale is required for commensurability: it has to be that one of the items is better than the other or else they are equally good. If neither is the case, they are incommensurable.

thing as incommensurability.[4] Like Joseph Raz (1986), I take the comparison with "some third thing" as a test, or "mark" of incommensurability (Raz 1986: 326).

> A test of incommensurability between two options [given that it is known that neither is better than the other], which yields a sufficient but not necessary condition of incommensurability, is that there is, or could be, another option that is better than one but is not better than the other. (Raz 1986, abstract of ch. 13, "Incommensurability")

The Continuum Argument is meant to establish that for any population, however large, that consists of people with excellent lives, there is a better population in which everyone has a drab life, barely worth living. It proceeds by constructing a finite sequence of populations of rapidly increasing size, with a slowly decreasing quality of life, the same for all individuals in a given population. As one moves along the sequence, the lives of the individuals get worse and worse, but they remain worth living. The argument makes an assumption, which is not questioned by Parfit, that:

> Adding people with lives worth living always makes the world better, if the added lives are of the same quality as those of all the others.[5]

Indeed, Parfit accepts what he calls the Simple View, according to which the marginal value of such additions not only always is positive, but it is never diminishing (Parfit 2016, p. 112).

Let´s go back to the Continuum Argument. The first population in the sequence is large (as large as one pleases, according to the argument) and consists of people with excellent lives. Each succeeding population is much larger than the preceding

[4] I think Parfit would agree. In the quote above, in order to establish that imprecise equality need not be transitive, he points out that "some third thing", Z, could be better or worse than only one of the imprecisely equal X and Y, but neither better nor worse than the other. He obviously implicitly assumes that Z would then be imprecisely equally as good as the other item. Otherwise the existence of such Z would be irrelevant to the transitivity issue. But if Z is imprecisely equally as good as Y, this implies that imprecise equality is not transitive, since Y is imprecisely equally as good as X but Z is not imprecisely equally as good as X. (It is either better or worse than X.) Parfit´s implicit assumption, which otherwise could be questionable, is necessarily satisfied if imprecise equality is nothing but incommensurability. It holds for all mutually incommensurable X and Y, and for all Z, that if Z is better or worse than, say, X, but neither better nor worse than Y, then Y cannot be precisely as good as Z. For then Y would have been better or worse than X (by the transitivity of these relations across precise equal goodness), contrary to the hypothesis. This means that Z must be incommensurable with Y: neither better nor worse than Y, nor (precisely) equally as good.

[5] Personally, I am not at all sure this view is correct. It is arguable that such additions might be axiologically neutral: they need not make the world either better or worse (see Rabinowicz 2009a). Here, however, I am willing to go along with Parfit and take this assumption as given.

one, but the lives of its members are slightly worse. The large increase in size is meant to compensate at each step for a small decrease in life quality: Each population is thus supposed to be better than the preceding one. The last, huge population consists of people with drab lives, of the muzak-and-potatoes variety, still worth living but only barely so. By the transitivity of betterness, this last population is better than the first one.

For Parfit, this is a repugnant conclusion. Indeed, he believes that a sufficiently large population of people with excellent lives is better than any population, however large, of people with drab lives.

To block the argument, Parfit suggests that, at some points in the population sequence, we will encounter an incommensurability (an 'imprecise equality'): the next population will be incommensurable with the immediately preceding one, instead of being better, as the Continuum Argument requires.

However, as we shall see, blocking the argument in this way carries considerable costs. The kind of incommensurability Parfit needs to posit has to be very thoroughgoing: It must be persistent in the sense to be explained. As we shall see, this persistency, while highly atypical, can be accounted for if incommensurability is interpreted, following Rabinowicz (2008, 2012), on the lines of the fitting-attitudes analysis, as permissibility of divergent attitudes regarding the compared populations. But even if the worry about persistency can be allayed, the conclusion will be that some of Parfit´s substantive value assumptions – in particular, his Simple View regarding the marginal value of increases in population size – need to be rejected.

2. Preparing the ground – informal discussion

The following principle expresses the key idea behind the Continuum Argument:

> Compensation: Small losses in quality can always be compensated, and indeed outweighed, by sufficiently large increases in quantity.
>
> Or, somewhat more precisely: For all positive levels of quality, if the move from one level to the other only involves a small quality loss, then there is an increase in quantity that would outweigh this loss.

In the Continuum Argument, as one moves forward in the population sequence, small losses in life quality are supposed to be outweighed, at each step, by increases in quantity – in the population size. Each population in the sequence is therefore supposed to be better than the immediately preceding one.

On the other hand, the argument´s conclusion contravenes the intuition that

> Large losses in quality cannot always be compensated by increases in quantity.

There are positive levels of quality such that the loss involved in the move from one level to another cannot be compensated by any increase in quantity, however large. This intuition lies behind the claim that the starting-point of the population sequence is better than its end-point.

In particular, large quality losses cannot be compensated by increases in quantity in cases of *weak superiority*, i.e., when there is some quantity of the good of a higher quality that is better than any quantity of the good of a lower quality. In the case Parfit focuses on, the higher good is an excellent life, while the lower good is a drab life.

'Weak superiority' as the label for this relationship between types of good was introduced in Arrhenius and Rabinowicz (2005, 2015). Other names that have been in use for the same relation are 'discontinuity in value' (Griffin 1986), 'radical superiority' (Handfield and Rabinowicz 2017) and 'lexical superiority' (Parfit 2016). Parfit´s label – 'lexical superiority' – is potentially misleading, which is why I here abstain from using it. To say that a type of good is lexically superior to another type of good, might suggest that *any* quantity of the former would be better than any quantity of the latter. In Arrhenius and Rabinowicz (2015), the latter relation is referred to as *strong superiority*. Griffin (1986) calls it 'trumping'. Strong superiority is not at issue in the case we consider: Parfit doesn't suggest that even a single excellent life would be better than any number of drab lives. Indeed, I think he would balk at such a claim. His perfectionism doesn't go that far.

To block the Continuum Argument we have three options:

> (1) deny that it is possible to construct the required sequence of types of lives;
>
> (2) deny the transitivity of betterness for the sequence of populations;
>
> (3) deny that small quality losses can always be compensated by increases in quantity.

Option (1): One might question whether the large gap between an excellent life and a drab life can be scaled in a finite number of small decreases in life quality. Small relative to what? What counts as a small decrease diminishes in absolute size as the

quality of a life goes down. But then we might never reach a drab life, barely worth living, from an excellent life, in a finite number of small steps.

This point can also be put in another way: It might be that for every population of people with lives of any given quality there is *some* decrease in this quality that can be outweighed by a sufficient increase in quantity – in the population size. But if we start with a population at a high quality level and then at each step decrease the quality by some amount that we can simultaneously outweigh by an increase in quantity, we might in this way go on for ever and still never be able to reach a population of low quality lives.[6]

This objection is plausible but not absolutely compelling. Defenders of the Continuum Argument might respond that even if decreases in life quality, as one moves along the life sequence, must be regarded as small at each step, they might still get us all the way down to a life that is barely worth living. Suppose, for example, that each decrease in quality, from one life to the next one, is of the same absolute size but this size is chosen in such a as to be counted as small even in the last step, at a point when the quality of a life is already very low. Then such decreases will be counted as being even smaller in all the previous steps. This construction would guarantee that a drab life can be reached from an excellent life in a finite number of small steps of equal size, provided that the quality distance between these two types of life is not infinitely large. Admittedly, this response makes strong assumptions about measurability of life quality: Unless the latter can be measured on an interval scale, it is meaningless to talk about equal-sized quality decreases.[7] But still, such a

---

[6] More generally, assume that

(i) For every life quality L and for every quantity k, there is some lower life quality L^- and (higher) quantity k^+ such that a population of k^+ L^--lives is better than a population of k L-lives.

From this it does *not* follow that

(ii) For every life quality L and quantity k and for every quality L′ lower than L, but still positive, there is some quantity k′ such that a population of k′ L′-lives is better than a population of k L-lives.

Assumption (i) allows us to construct a sequence of populations of decreasing quality and increasing quantity in which every element is better than its immediate predecessor: kL, k^+L^-, $k^{++}L^{--}$, $k^{+++}L^{---}$, ... But even if we extend this sequence to infinity, all qualities in the sequence might be higher than L′. Indeed, even the quality to which this sequence converges in the limit might be higher than L′.

[7] This measurability assumption is especially problematic in view of the fact that for Parfit life quality is not reducible to wellbeing level. While the latter is a component in life quality, life quality is a broader concept. When we consider the Repugnant Conclusion, reducing the quality of a life to its wellbeing level would have unwelcome consequences. Parfit allows that a drab life might have the same overall level of wellbeing as, say, a "roller-coaster" life in which highs alternate with lows, if in in such a life the highs only barely outweigh the lows. In Parfit´s view, the Repugnant Conclusion is truly repugnant only when it is a claim about drab lives –which "contain nothing bad, but very little that was good." What´s repugnant is that a population of drab lives could be better than a population of excellent lives. An analogous claim about roller-coaster lives or about happy lives of extremely short duration (stretching over "one happy day, or one ecstatic hour") would not be (so) repugnant (Parfit 2016, p. 118). He emphasizes that "there are important differences between the quality of people´s lives and the amount of well-being per person" (ibid.). He apparently allows that two lives can have the same "amount of

reply cannot be dismissed out of hand. To the extent that we do have intuitions about life quality losses or gains being small or large as compared with others, we do seem to be prepared to measure such losses and gains on something like a (rough) interval scale.

Option (2): One might want to deny the transitivity of betterness. Larry Temkin, who champions this proposal, suggests that, in all-things-considered value comparisons, we should interpret betterness as an "essentially comparative" concept (see, for example Temkin 2012). Therefore, in comparing alternative outcomes, we should not start with assessing how good or bad each outcome is in itself. We should not evaluate outcomes in isolation, but only relative to each other. Given this Essentially Comparative View, there is no guarantee that betterness is going to be transitive. Temkin contrasts this perspective with the Internal Aspects View, according to which value comparisons are secondary to intrinsic evaluations of outcomes. On that view, the transitivity of betterness is guaranteed. (Cf. Temkin 2012, sections 7.6, 7.7.) But then, if it is the essentially comparative conception that should be used in evaluating outcomes, such as alternative populations, we can block the Continuum Argument by denying transitivity. Indeed, we can then accept that the argument´s population sequence forms a betterness cycle: Each population in the sequence is better than its immediate predecessor, but the last population is worse than the one we have started with.

In a posthumously published paper (Parfit 2017), Parfit willingly admits that betterness can be understood in these different ways, but he suggests that (something like) the Internal Aspects View still is appropriate when it comes to evaluating and comparing alternative outcomes:

> Temkin rightly assumes that the phrase "better than" can be used in different senses, which refer to different relations. Temkin argues that, though some outcomes are intrinsically better than others in a sense that is transitive, such claims may apply to only a "severely limited part of the normative realm." I believe that many outcomes, and many other things, are intrinsically good or bad, in ways that make them better or worse than others. Such goodness or badness is not essentially comparative, since it does not consist in being related in certain ways to other things. (Parfit 2017, p. 139)

Option (3): This is Parfit´s preferred alternative. The other two options are discarded; he assumes that the relation of betterness between populations is transitive

wellbeing", but still the quality of one of them might be higher. He also seems to allow that the the quality of one life (of the roller-coaster type) might be higher than that of another life (of the muzak-and potatoes variety) even though the average wellbeing is the same in both.

and that it is possible to construct a finite slowly decreasing life sequence that leads from an excellent life to a drab life. What he instead denies is that small quality losses can always be compensated by increases in quantity. At some points in the Continuum Argument's population sequence X_1, ..., X_n, we will encounter an incommensurability according to Parfit: For some $i < n$, the next population X_{i+1} will not be better than the immediately preceding X_i, as is required by the Continuum Argument. Instead, the two populations will be incommensurable with each other. This makes it possible to reject the repugnant implication that X_n must be better than X_1, and indeed to insist on the opposite being the case: X_1, a population of people with excellent lives, is better than X_n, a population of people with drab lives, barely worth living, however large the latter population might be.

As an aside, it should be pointed out that bringing incommensurability in at just one point in the population sequence, between, say, X_i and X_{i+1}, is not a viable option. If for every $j < n$ distinct from i, X_{j+1} were better than X_j, the transitivity of betterness would imply that (i) X_i is better than X_1 and (ii) X_n is better than X_{i+1}. Since Parfit claims that (iii) X_1 is better than X_n, it follows from (i), (iii) and (ii), again by the transitivity of betterness, that X_i is better than X_{i+1}, contrary to the assumption that that these two populations are incommensurable.

As was shown in Handfield (2014), this inconsistency can be avoided if we postulate that there are *at least two* points of incommensurability in the sequence.[8] And indeed, Parfit seems to think that incommensurability ("imprecise equality") might well come in at *every* point in the population sequence:

> [Continuum] arguments assume that [...] any slight loss of quality could be outweighed by a sufficient gain in quantity. If we assumed precision, it would be hard to reject these arguments. We would have to claim that *any* slight loss of quality would outweigh any gain in quantity.[9] As several writers claim, that would be very implausible. Compared with the existence of some number of people, it would not always be worse if instead there existed many more people who would have a slightly lower quality of life. But we should deny that such truths would be

[8] A simple example shows that two such points can suffice for a sequence of an arbitrary length $n > 2$: Suppose that, for some i such that $1 < i < n$, X_i is incommensurable both with X_{i-1} and with X_{i+1}. With these two exceptions, there are no other points in the sequence at which incommensurability intervenes between adjacent populations: For all other adjacent pairs, the second population in the pair is better than the first, as in the Continuum argument. But still the first population in the sequence is better than the last one. By the transitivity of betterness it then follows that X_{i-1} is better than X_{i+1}, but this implication does not create any inconsistency. To complete the construction, assume that X_i is incommensurable not only with its immediate neighbors but also with all the other populations in the sequence. (This makes the example rather unrealistic in case of long sequences, but more realistic examples could perhaps be constructed if needed.)

[9] This is a *non-sequitur*. See my comment below.

> precise. We should then claim that *no* slight loss in quality would either be outweighed by, or outweigh, any such gain in quantity. It would not be better if there existed many more people whose quality of life would all be lower, since two such worlds would at most be imprecisely equally good. (Parfit 2016, p. 120, my emphasis)

Parfit´s suggestion that *no* slight loss in quality can be outweighed by any gain in quantity seems unnecessarily radical. It is not needed to block the Continuum argument. As I pointed out above, it is enough if incommensurability intervenes at just two points in the population sequence. At other points, the next population may well be better than its immediate predecessor.[10]

But let´s put this matter aside. Parfit´s central suggestion, I take it, is that at least some adjacent populations in the Continuum Argument´s population sequence, will be mutually incommensurable. To assess this proposal, we first need to develop a simple formal framework. This will be done in the next section.

3. Preparing the ground – formal framework

The domain we going to consider consists of populations in which everyone´s life is worth living and which are homogeneous in life quality – everyone has the same type of life. We refer to types of life as L, L′, ..., and to quantities – population sizes – as k, k′, ..., etc. Thus, in what follows, kL will stand for a population of k people with lives of type L. We do not distinguish between populations on the basis of the personal identity of its members. This is not needed for the discussion that follows.

For any L such that the domain contains some L-population, we take it that any quantity of L is possible: If he domain contains a population kL for some k, then it contains populations kL for every k. There is no upper limit on population sizes.

There are two kinds of betterness relations we need to consider: one between types of life and the other between populations. To distinguish between them we use the subscript T for the former relation. Thus, $L \succ_T L'$ stands for the claim that L is a better type of life than L′, while $kL \succ k'L'$ states that a population of k people with lives of type L is better than a population of k′ people with lives of type L′.

[10] Parfit might have been driven to this radical suggestion by his mistaken assumption that, given precision, i.e., given the absence of incommensurabilities, critics of the Continuum Argument "would have to argue that *any* slight loss of quality would outweigh any gain in quantity" (see the quote above, my emphasis). That is, the critics′ claim would have to be, according to Parfit, that every step in the population sequence is a worsening. This assumption is incorrect. Clearly, it would be enough, given precision, if only some steps in the sequence were worsenings. This would be sufficient for the Continuum Argument to crumble.

We assume that, ceteris paribus, i.e., keeping the size of population constant, lives of a better type make better populations:

Principle of Quality: For all L, L′ and all quantities k, $kL \succ kL'$ iff $L \succ_T L'$.

Thus, the two relations are closely connected.

Indeed, the relation between types can be defined in terms of the relation between populations:

$L \succ_T L' =_{df} 1L \succ 1L'$.

In other words, a type of life L is better than another type of life L′ whenever a life of type L is better than a life of type L′.[11] Therefore, in what follows, we shall freely move from betterness claims regarding types of lives to betterness claims regarding lives.

Given this definition, Principle of Quality follows if we impose the following constraint on $\succ$:

Quality Constraint: For all L, L′, k and k′, if $kL \succ kL$, then $k'L \succ k'L'$.

And conversely, the Principle of Quality in its turn implies both this Quality Constraint and the above definition.

Along with the Principle of Quality, I am going to assume the corresponding Principle of Quantity. Since we only consider populations of people with lives worth living, I assume, following Parfit, that larger populations are better, if life quality is kept constant:

Principle of Quantity: For all k, k′ and L, $kL \succ k'L$ iff $k > k'$.

Here, > stands for the relation "greater than" between quantities.

As for the relation of betterness among populations, I assume that this relation is asymmetric and transitive.

[11] Analogously, L an L′ can be defined as being equally good if and only if 1L is equally as good as 1L′, and as incommensurable if and only if 1L is incommensurable with 1L′. It is often assumed that wellbeing levels form a linear ordering: if two such levels are distinct, then one is higher than the other. But, as we have seen above, the type (= quality) of a life is not reducible to its wellbeing level. Thus, it should be possible for distinct L and L′ to be equally good or even incommensurable.

Asymmetry: If $kL \succ k'L'$, then $k'L' \not\succ kL$.

Transitivity: If $kL \succ k'L'$ and $k'L' \succ k"L"$, then $kL \succ k"L"$.[12]

Indeed, I shall also assume that betterness and worseness (= the converse of betterness) are *transitive across equal goodness*, i.e. that whatever is better (worse) than kL must be better (worse) than $k'L'$ if kL and $k'L'$ are (precisely) equally good. Thus, letting $\approx$ stand for the transitive, symmetric and reflexive relation of equal goodness, I assume, as part of the condition of Transitivity, that

(i) if $k"L" \succ kL$ and $kL \approx k'L'$, then $k"L" \succ k'L'$
and

(ii) if $kL \succ k"L"$ and $kL \approx k'L'$, then $k'L' \succ k"L"$.[13]

According to Parfit, an excellent life is weakly superior ("lexically superior", as he puts it) to a drab life. The relation of weak superiority between types of life can easily be defined in our framework:

L is *weakly superior* to $L' =_{df}$ For some k, $kL \succ k'L'$ for all k'.

That is, L is weakly superior to L' if some (sufficiently large) quantity of L is better than every quantity of L'.
Let me now introduce another concept that will be useful in what follows.

L is *exchangeable* for $L' =_{df}$ For every k, there is some k' such that $k'L' \succ kL$.

That is, L exchangeable for L' if for every quantity of L there is a quantity of L' that is better.[14]

---

[12] By the Principle of Quality, this implies that the betterness relation between types of life is also asymmetric and transitive. Note, by the way, that the transitivity of the latter relation has not been questioned in the discussion of the Continuum Argument. It is the transitivity of the betterness relation between populations that has been a matter of dispute.

13 A much more economical representation that entails all these standard conditions on $\succ$ and $\approx$ starts from a relation $\succcurlyeq$ of being <u>at least as good</u>. This relation is assumed to be reflexive and transitive. Betterness and equal goodness as then defined as the asymmetric and the symmetric parts of $\succcurlyeq$, respectively:

$kL \succ k'L' =_{df} kL \succcurlyeq k'L'$ but not $k'L \succcurlyeq kL$; $kL \approx k'L' =_{df} kL \succcurlyeq k'L'$ and $k'L' \succcurlyeq k'L'$.

[14] Note that exchangeability, as here defined, is a non-symmetric relation (which might not be how we use this term in ordinary language). If L is exchangeable for L', it doesn't follow that L' is exchangeable

As we have seen, the main idea behind the Continuum argument is Compensation: Small losses in quality can always be outweighed by sufficiently large increases in quantity. For this outweighing to be possible, the type of life of a higher quality must be exchangeable for the type of a lower quality.

Now, suppose there is a sequence of types of life, L_1, L_2, ... , L_n, such that L_1 is weakly superior to L_n.Thus, for some quantity k_1, k_1L_1 is better than any quantity of L_n. (Think of L_1 as an excellent life and of L_n as a drab life.) Then it *cannot* be true that every type L_i in this sequence is exchangeable for its successor L_{i+1}. Otherwise, it would be possible to construct a sequence of populations starting with k_1L_1

$$k_1L_1, k_2L_2, ..., k_nL_n$$

in which each population is better than the immediately preceding one. By the transitivity of betterness, it would then follow that k_nL_n is better than k_1L_1. Which, given the asymmetry of betterness, is incompatible with k_1L_1 being better than any quantity of L_n.

But if L_i is not exchangeable for L_{i+1}, how can these types be related instead?

One possibility is that L_i is weakly superior to L_{i+1}.[15] But how can this be if we suppose that at each step in the sequence the next type is only slightly worse than its immediate predecessor?

To explain this, we might appeal to the idea of *the diminishing marginal value of quantity*. It might be that k_iL_i is better than every quantity of L_{i+1} because increases in the quantity of L_{i+1} steeply diminish in marginal value. As a result, no quantity of L_{i+1} can ever reach the value of k_iL_i, even though L_i is only slightly better than L_{i+1}. (Cf. Arrhenius & Rabinowicz 2005, 2015. See also Jensen 2008.)

Parfit (2016, p. 112) finds this option implausible; he rejects the Diminishing Value View. The argument he uses appeals to the analogy between goodness and badness:

> The existence of [...] wretched people would not have a badness that would diminish as the number of such people grew, so that it mattered less and less whether more such people exist. The badness of more such suffering would never decline. [....] we cannot plausibly either apply some Diminishing Value View to lives that are bad, or restrict this view to lives that are good.

---

for L.

[15] Indeed, it is the only available possibility as long as the betterness relation between populations is supposed to be *complete*, i.e., as long as we do not allow for incommensurabilities. In Arrhenius and Rabinowicz (2005. 2015) it is shown (using a somewhat different formal framework) that given this completeness condition on $\succ$, any sequence L_1, ..., L_n such that L_1 is weakly superior to L_n must contain some L_i that is weakly superior to its immediate successor in the sequence.

But is it a convincing argument? Why can ' t we restrict the Diminishing Value View to good lives only, i.e. accept diminishing value but reject diminishing *dis*value? Indeed, Parfit himself admits that "there are some asymmetries between suffering and happiness, and some of the other things that can make lives good or bad" (ibid.). So why cannot there be an asymmetry between goodness and badness in this respect as well?

Nevertheless, Parfit is not willing to entertain this possibility. He accepts what he calls the *Simple View*:

> the Simple View claims: Anyone´s existence is in itself good if this person´s life is worth living. Such goodness has non-diminishing value, so if there were more such people, the combined goodness of their existence would have no upper limit. (ibid.).[16]

This leads him to reject the suggestion that at some point in the slowly decreasing life sequence L_1, ..., L_n, we could encounter a type of life, L_i, that is weakly superior to the next type L_{i+1}, in spite of it being only slightly better than the latter.

As the Simple View is literally stated, it only denies that the marginal value of added lives can diminish. But what about the possibility that this marginal value could sometimes *increase*? For example, that the value of a population could radically increase upon reaching a certain size, perhaps even increase to infinity? I believe, however, that Parfit would reject this possibility as well. He seems to disallow any holistic value effects of increases in population size. On his view, if I understand him correctly, the value added by a life to a population is simply its intrinsic value. Thus, if the added life does not affect the life quality of other population members, it can never add more value to the population, or less value, than what it is intrinsically worth. Therefore, I take it that the Simple View should be understood as the claim that increases in the population size have a *constant* marginal value. I will come back to the Simple View later, in the last section, but right now let us simply take it as given that Parfit is not prepared to allow that a type of life can be weakly superior to another type of life that is only slightly worse.

But then, if none of the types of life in the gradually decreasing type sequence leading from an excellent life to a drab life can be exchangeable for its immediate successor, as this would lead to the Repugnant Conclusion, and if none of these types can be weakly superior to its immediate successor, what else is left?

[16] Note though that earlier in the paper Parfit provides a different, much weaker formulation of the Simple View: "the *Simple View*: Anyone's existence is in itself good, and makes the world in one way better, if this person's life is good to live, or worth living." (Parfit 2016, p. 110)

4. Enter incommensurability

What is left is the possibility that for some pairs L_i and L_{i+1} of adjacent types in this gradually decreasing type sequence stretching from an excellent life to a life barely worth living, the relation between L_i and L_{i+1} admits *incommensurability*. And indeed, admits incommensurability of a very thoroughgoing kind, which I will call *persistent* incommensurability. By this I mean that some quantity k_i of L_i is such that L_{i+1} is worse than k_iL_i in smaller quantities and incommensurable with k_iL_i in *all* larger quantities. Thus, if k_iL_i isn't better than $k_{i+1}L_{i+1}$, then these two populations are incommensurable and this incommensurability would persist however much the quantity of L_{i+1} were increased. This would mean that we cannot repair the Continuum Argument by increasing the quantity of L_{i+1}. No quantity of L_{i+1} would do the job required by the Continuum Argument, since no such quantity is better than k_iL_i.

Here´s the definition of this relation between types of life:

> L admits *persistent incommensurability* with L′ $=_{df}$ Some quantity k of L is incommensurable with every quantity of L′ at least as large as some k′.

This quantity k of L will then be better than every quantity of L′ smaller than the threshold k′.[17] Thus, no quantity of L′ will be better than kL. Which means that L is not exchangeable for L′.

As for weak superiority, if L admits persistent incommensurability with L′, this certainly does not imply that L is weakly superior to L′, but it does not exclude it either: Even if kL is incommensurable with every sufficiently large quantity of L′, there might be some quantity of L larger than k that is better than any quantity of L′. We can however exclude this possibility if we stipulate that persistent incommensurability with L′ extends to all quantities of L at least as large as k. We can define this more demanding relation as follows:

> L admits *strictly persistent incommensurability* with L′ $=_{df}$ For every quantity of L at least as large as some k, this quantity of L is incommensurable with every quantity of the other type L′ at least as large as some k′.

[17] Obviously, kL cannot be worse than or equally as good as some quantity of L′ that is smaller than this incommensurability threshold k′. For then, by the Principle of Quantity and by Transitivity, kL would have to be worse than k′L′, contrary to the hypothesis.

But does the incommensurability between two adjacent types of life have to be so thoroughgoing, if we want to block the Continuum argument by an appeal to incommensurability? The answer is yes. The following can be proved:

> Trilemma: For any two types of life L and L′, exactly one of the following three relations must obtain: (i) L is exchangeable for L′, (ii) L is weakly superior to L′, or (iii) L admits strictly persistent incommensurability with L′.

The proof of the Trilemma (see Appendix) relies on Transitivity and the Principles of Quality and Quantity.[18]

Thus, to block the Continuum Argument by an appeal to incommensurability, Parfit needs to postulate a very radical form of incommensurability.

Let′s consider for a moment the dialectics of a hypothetical debate between Parfit and an adherent of the Continuum Argument.[19] Parfit, who wants to insist that an excellent life (L_1) is weakly superior to a drab life (L_n), might be asked to specify a number k_1 of L_1-lives that in his view would be better than any number k_n of L_n-lives. If he does come up with a definite answer, which might not be that easy, the adherent of the Continuum Argument needs to counter this proposal. Thus, she needs to specify an appropriate population sequence: a sequence of (what she claims) better-and-better populations of people with worse-and-worse lives which starts with k_1L_1 and ends with k_nL_n. k_n may be as large as she pleases. To come up with a plausible sequence of this kind might not be easy. But, if this demand is met, then it is up to Parfit to try to find a fault in his opponent′s proposal: He needs to point to some specific types of life in the offered sequence that instead of being exchangeable for their successors, as the opponent claims, admit strictly persistent incommensurability with the types that immediately follow.[20] This might not be easy. But even if Parfit would do it, it would not be the end of the matter. The pro-

[18] In Handfield & Rabinowicz 2017, a very similar trilemma, though formulated in somewhat different terminology and with weak inferiority replacing weak superiority, was proved for bad types more specifically, for types of harm. Here, I prove it for good types – types of life worth living.

[19] I am indebted to John Broome for making me think harder about this issue.

[20] Alan Hájek has questioned this (in private communication). Why should there be this requirement on Parfit? To give an analogy, a critic of the *sorites* need not point to the specific number of grains of sand at which heaps begin. I think though that there is an important dialectical difference between these two cases: While everyone agrees that the conclusion of the *sorites* is absurd, the Continuum Argument has its adherents. The critic of the *sorites* is therefore in a stronger dialectical position. He need not be as specific in his critique as the critic of the Continuum Argument. Still, while imperfect, this analogy is instructive. In both cases, it might be *indeterminate* where exactly the relevant break point in the sequence is located. In particular, it might be indeterminate which types of lives is the Continuum Argument's life sequence admit of strictly persistent incommensurability with their immediate successors. I do not consider the possibility of indeterminacy in this paper, but the general issue of how indeterminacy, or vagueness, can interact with incommensurability is discussed in Rabinowicz (2009b).

ponent of the Continuum Argument might come back and offer *another* population sequence (starting with k_1L_1 and continuing all the way to k_nL_n, for an appropriately large k_n), which Parfit in his turn would then again need to find a fault in, and so on. It is not clear how this debate could ever be conclusively resolved. Note that both sides have burdens of proof, as both come up with positive – but opposing – claims.

In what follows, I gloss over the complexity of this dialectics. Instead, I focus on the very concept of persistent incommensurability. What I am going to say will also apply to the stronger notion of strictly persistent incommensurability.

5. Assessing the costs

Persistent incommensurability is a strange animal. In the standard cases used in the literature to illustrate incommensurability, on one of the compared items is better than the other in some relevant respects and worse in some others. The same applies to the case we focus on, in which the compared items are (homogeneous) populartions and the relevant respects, or dimensions of comparison, are quality and quantity: type life and population size. However, in the standard cases of multidimensional comparisons, if two items are incommensurable, this incommensurability is not supposed to be persistent: A sufficiently large improvement (worsening) of one of the items in a relevant dimension will eventually make it better (worse) than the other item.

In the well-known *small-improvement argument* for incommensurability, one starts with two items, X and Y, none of which is better than the other. The aim is to prove that they are not (precisely) equally good either and they therefore must be incommensurable. To achieve this aim, one considers a small improvement of one of the items, say Y, in some relevant respect. This slightly improved variant of Y, Y^+, is better than Y, but it still is not better than X. This shows that X and Y cannot be equally good, since betterness is transitive across equal goodness.

It is significant that this argument involves a *small* improvement. Given that X isn't better than Y, it would strain credulity to think that *no* improvement of Y in the relevant respect, however large, could make the improved variant of Y better than X. Therefore, if k_iL_i isn't better than $k_{i+1}L_{i+1}$, it might strain credulity to think that no increase in the quantity of L_{i+1}, however large, could make this quantity of L_{i+1} better than k_iL_i. This is especially worrisome if, as the Simple View has it, the marginal value of added lives never diminishes. But even if this value is allowed to diminish, it's still possible that large increases in quantity can considerably increase the value of L_{i+1}-population. Thus, to put it cautiously, it seems that persistent incommen-

surability which Parfit needs to block the Continuum Argument is highly atypical. This puts his proposal under stress.[21]

Let me develop this line of thought a little bit further. Ruth Chang has made an influential suggestion that in cases of multidimensional comparisons, if neither item is better than the other and they are not equally good, the two items can be assumed to be *on a par*. They are mutually incommensurable (in the sense in which I use the term), but still in the same league, so to speak.[22] They are not incomparable: for Chang, parity is a form of comparability. It is plausible to suppose, I think, that when Parfit's "imprecise equality" is just another name for parity. Now, it is often taken to be a feature of parity, as opposed to outright incomparability, that if two items, X and Y, are on a par, and it is possible to considerably improve one of them, say, Y, in one or several relevant respects, then this considerately improved version of Y will be better than X. In other words, parity isn't supposed to be persistent.

Indeed, Qizilbash (2018) takes such impersistency to be "the mark of parity". On his interpretation of that relation, parity is what he calls "rough equality":

> On this view [= the rough equality view], when parity holds between items [...] while some slight change in value may not tilt the balance in favour of one of them, any significant increase in the value of one of the items will make it better, and any significant reduction in the value of one will make it worse, than the other. This feature of parity is the 'mark of parity' on the rough equality view. [...] this mark [...] is implied by a central component of the rough equality view: the assumption that while parity is a form of equality and a distinct relation, it is not precise equality. The mark of parity also distinguishes parity from 'incomparability'. If there were 'incomparability', on this view, while one is not better than the other even some significant increase (or reduction) in the value of one option would not make it better (worse) than the other. (Qizilbash 2018, p. 90)

[21] This worry about persistency of incommensurability was first raised in Handfield and Rabinowicz (2017), in connection with the Spectrum Argument against the transitivity of betterness. (In that paper, strictly persistent incommensurability was referred to as "radical" incommensurability.) To block the Spectrum Argument by an appeal to incommensurabilities also requires these incommensurabilities to be strictly persistent. Which is not surprising since that argument is a close cousin of the Continuum Argument. In the next section, I am going to suggest how the worry about persistency could be put to rest. Thus, the present paper might be seen as an attempt to solve the problem posed in my earlier joint work with Toby Handfield.

[22] "Parity typically holds between items that bear very different aspects of V [V refers to the covering value, or covering consideration, with respect to which the items are being compared; aspects of V are different relevant respects] and yet are nevertheless 'in the same neighbourhood' of V-ness." (Chang 2016, p. 193) Different "aspects of V" are different dimensions of V – different relevant respects of comparison.

He contrasts this rough-equality account of parity with the one I myself have put forward in Rabinowicz (2008, 2012) and suggests that the latter account fails to entail that parity must have this mark of impersistency. He is right, as I am now going to show. My account takes its departure from the *fitting-attitudes analysis* of value but interprets fittingness as permissibility rather than requiredness, i.e., it allows that there sometimes might exist a range of different fitting (permissible) attitudes towards an item. Here is a short sketch of my proposal:

Value relations are analysed in terms of the class, K, of permissible preference orderings of the domain of items under consideration. An item X is *better* than another item Y if and only if it ought to be preferred to Y, i.e., if and only if every permissible preference ordering (= every ordering in K) ranks X above Y. They are *equally good* if and only if every K-ordering ranks them equally. Or, what amounts to the same, if and only if one ought to be indifferent between X and Y. Incommensurability arises when it is permissible for preferences to diverge. In particular, X and Y are *on a par* if and only if some orderings in K rank X above Y, but some other K-orderings rank Y above X. That is, they are on a par if and only if it is permissible to prefer X to Y but also permissible to have the opposite preference.

The orderings in K might, or might not, contain gaps in some places. X and Y are *incomparable* if and only if every permissible ordering contains a gap as far as the comparison between these two items is concerned. That is, if and only if it is impermissible to prefer one of them to the other or to be indifferent. Needless to say, this account makes incomeparability a rare phenomenon.

Orderings in K are meant to represent permissible *overall* preferences, i.e. preferences that take into account all the relevant respects of comparison. Parity will typically arise in cases of multidimensional comparisons, if one item, X, ranks higher than the other, Y, in some respects and lower in other respects. The overall preference as regards these items will then depend on how the different respects (dimensions) are weighted against each other. It is unrealistic to expect that there is a unique correct assignment of weights to dimensions. Different weight assignments will normally be admissible. Therefore, it might well be the case that an overall preference for X over Y and an overall preference for Y over X might both be permissible – one based on one admissible assignment of weights to the relevant dimensions and the other based on another admissible weight assignment. As a result, X and Y will be on a par.

Now, Qizilbash suggests that this account, unlike his own, does not *entail* that parity cannot be persistent. He is right, of course. However, even on my proposal, parity typically is not going to persist if one of the items that are on a par gets more-and-more improved in some relevant respect. Sooner or later, when the improve-

ment becomes large enough, the preference for the other item will typically stop being permissible, which will put an end to parity.[23]

Nevertheless, it is an important feature of my account that it at least in principle allows for persistent (and indeed strictly persistent) parity. I will demonstrate how this possibility can arise in the next section.

6. Accounting for persistence

First, let me sketch how my approach applies to value comparisons between populations.

The domain of items we consider is, as previously, a set of possible populations with different sizes k and different types of life, L, the same for every person in a population and worth living. There is no upper limit on size. K is the class of permissible preference orderings of that domain. Every such ordering P of populations induces the associated ordering P_T of types of life by the definition: P_T ranks L higher than L′ $=_{df}$ P ranks 1L higher than 1L′. We take it that all K-orderings induce the same ordering of types of life: for all such orderings, P and Q, $P_T = Q_T$.[24] We also assume that K-orderings satisfy the preferential versions of the conditions previously imposed on the relation of betterness between populations. Thus, permissible preferences are asymmetric and transitive (also across indifferences) and they satisfy the preferential versions of the Principles of Quality and Quantity:

> *Preferential Principle of Quality*: For all K-orderings P, all k, L and L′, P ranks kL above kL′ iff PT ranks L above L′.[25]
>
> Preferential Principle of Quantity: For all K-orderings P, all L, k and k′, P ranks kL above k′L iff k > k′.

[23] Impersistence of parity is also to be expected on another recent account of that relation, due to Ursula Andreou (Andreou 2015). On that account, two items are on a par with respect to some covering value V if and only if they belong to the same category (the same 'league') with respect to V, but nevertheless are mutually incommensurable with respect to V, i.e., neither is a better V nor are they equally good specimens of V. Categories of a given value are supposed to be linearly ordered from higher to lower: excellent specimens of V, good specimens, mediocre ones, etc. Unless the categories are very broad, if X and Y are on a par, a significantly improved (worsened) version of Y will belong to a higher (lower) category than X. Thus, it won't be on a par with X. (Indeed, it won't anymore be incommensurable with X.) Which means that even on this account parity typically isn't going to persist.

[24] In a more general framework, this assumption can easily be given up, though. Indeed, there are good reasons to give it up if we think of life types as being multi-dimensional. Rankings of life types can then be expected to differ from one permissible preference ordering to another.

[25] Just as the betterness ordering of types of life can be defined in terms of the betterness ordering of populations, the P_T-ordering of types of life is definable in terms of P: P_T ranks L above L′ if and only if P ranks 1L above 1L′.

Given my analysis of betterness and equal goodness in terms of permissible preferences, these conditions on K imply the corresponding conditions on betterness and equal goodness: betterness is asymmetric and transitive (also across equal goodness), and it satisfies the Principles of Quality and Quantity.

Let me now define two preferential relations that will be useful in the following discussion: the preferential variants of weak superiority and exchangeability. Suppose that P is an ordering of possible populations:

> L is *P-exchangeable* for L′ iff for every quantity of L there is a quantity of L′ that P ranks higher.

> L is *weakly P-superior* to L′ iff there is some quantity of L that P ranks higher than every quantity of L′.

It is now easy to show how strictly persistent incommensurability can arise in my model.

> Lemma: If K contains an ordering P such that L is weakly P-superior to L′ and another ordering Q such that L is Q-exchangeable for L′, then L admits strictly persistent incommensurability with L′, and indeed strictly persistent parity with L′.

> Proof: If L is weakly P-superior to L′, then for some k, (i) P ranks kL higher than every quantity of L′. But if L is Q-exchangeable for L′, then (ii) for some k′, Q ranks k′L′ higher than kL. By the properties of permissible orderings (more exactly, by the transitivity of permissible preferences and the Preferential Principle of Quantity), these implications can be extended to every quantity k^+ of L at least as large k: (iii) P ranks k^+L higher than every quantity of L; but (iv) for some k′, Q ranks higher than k^+L every quantity of L′ at least as large as k′. (iii) and (iv) together entail that L admits strictly persistent parity with L′.

It is not difficult in this model to provide an example of a sequence of types, L_1, ..., L_n, in which each type is better than but not weakly superior to its immediate successor, while L_1 is weakly superior to L_n. Clearly, if the latter holds, then it cannot be that every type in this sequence is exchangeable for its immediate successor. Therefore, by Trilemma, for some L_i in L_1, ..., L_n, L_i must admit strictly persistent incommensurability with L_{i+1}.

Indeed, there must be at least two types of life in this type sequence hat admit of

strictly persistent incommensurability with their immediate successors. If there was only one such type, then – by the Trilemma – all other types in the sequence would be exchangeable for their immediate successors. But then it would be possible, on the basis of this type sequence, to construct a sequence of populations in which incommensurability would appear at just one point while at every other point the next population would be better than its immediate predecessor. But we already know that this is impossible if betterness is transitive (see above, section 2).

To provide a model with the simplest possible type sequence of this kind, a sequence that consists of just three types, L_1, L_2, and L_3, suppose that for some permissible orderings P and Q,

both P_T and Q_T rank L_1 higher than L_2 and L_2 higher than L_3,

L_1 is weakly P-superior to L_2, while L_2 is P-exchangeable for L_3,

L_1 is Q-exchangeable for L_2, while L_2 is weakly Q-superior to L_3.

By the Lemma above, these conditions imply that L_1 admits strictly persistent parity with L_2 and L_2 admits strictly persistent parity with L_3. By the transitivity of permissible preferences, they also imply that L_1 is weakly P-superior to L_3 and that it likewise is weakly Q-superior to L_3.[26]

If we in addition suppose that orderings P and Q exhaust class K of permissible orderings, it will follow that L_1 is better than L_2 and L_2 is better than L_3, and that L_1 is weakly superior to L_3, just as required.[27]

It might be noted that nothing would change if we allowed for yet another permissible preference ordering, R, which is cautious and avoids adjudicating

[26] Proof: (i) Since L_1 is weakly P-superior to L_2, there is some k such that P ranks kL_1 above every quantity of L_2. Consider any quantity k' of L_3. Since P_T ranks L_2 higher than L_3, Preferential Principle of Quality implies that P ranks $k'L_2$ above $k'''L_3$. But then, by the transitivity of permissible preferences, P ranks kL_1 above $k'L_3$. Thus, L_1 is weakly P-superior to L_3. (ii) Since L_2 is weakly Q-superior to L_3, there is some k such that Q ranks kL_2 above every quantity of L_3. Since Q_T ranks L_1 higher than L_2, Preferential Principle of Quality implies that Q ranks kL_1 above kL_2. But then, again by the transitivity of permissible preferences, Q ranks kL_1 above every quantity of L_3. Thus, L_1 is weakly Q-superior to L_3.

[27] Proof: To prove that L_1 is weakly superior to L_3, we need to show that some quantity of the former is better than every quantity of the latter. Since L_1 is weakly P-superior to L_3, there is some k such that (i) P ranks kL_1 above every quantity of L_3. Similarly, since L_1 is weakly Q-superior to L_3, there is some k' such that (ii) Q ranks $k'L_1$ above every quantity of L_3. Now, either $k = k'$, or one of these two quantities is larger than the other. Without loss of generality, assume that $k \geq k'$. Then, by the transitivity of permissible preferences, (ii) implies that (iii) Q ranks kL_1 above every quantity of L_3. So, by (i) and (iii), both P and Q rank kL_1 above every quantity of L_3. Since P and Q exhaust class K, this implies that kL_1 is better than every quantity of L_3.

between P and Q. It can be defined as the intersection – the common part – of P and Q. Allowing for such cautious orderings seems reasonable. But if R is the common part of P and Q then adding it to K wouldn't affect any the conclusions we have reached above.

Thus, our modeling makes room for type sequences of the kind Parfit requires. In principle, we could have a type sequence which starts with an excellent life and ends with a drab life, in which no type of life is weakly superior to its immediate successor but the first type is weakly superior to the last. This becomes possible if we allow some pairs of adjacent types in the sequence to admit strictly persistent incommensurabilities.

7. Are we home then?

Not quite, I am afraid. While the abstract model described in the preceding section makes room for type sequences of the required kind, I don't think that there is a way to square this model with some of Parfit´s substantive value assumptions, and in particular with his Simple View. According to the latter, adding lives of the same type to a population has a non-diminishing marginal value. Parfit relies on this view when he denies that a type of life can be weakly superior to another type that is only slightly worse. In the slowly decreasing life type sequence that leads from an excellent life to a drab life, no type of life is supposed to be weakly superior to its immediate successor. But, in my modeling of such sequences, I still had to postulate the existence of permissible orderings in which some types are *preferentially* weakly superior to their immediate successors. (See the preceding section; a type is preferentially weakly superior to another type in an ordering P if and only if it is weakly P-superior to that type.) Can orderings of this kind be *permissible* if it is out of the question that the value of added lives could sometimes diminish? This may well be denied.[28]

However, Parfit's Simple View is highly questionable: It can be shown to have counter-intuitive implications. In a recent paper, Karsten Klint Jensen has proved the following result: If an excellent life is weakly superior to a drab life, as Parfit

[28] Actually, there seem to be two different ways of accounting for a type L being preferentially weakly superior to another type L′, despite it being only slightly better than L′. If L is weakly P-superior to L′, this might be either because the marginal impact of adding more and more L′-lives steeply decreases and converges to zero as the number of these lives goes to infinity, or because the L-population becomes radically upgraded in P-ranking upon it reaching a certain size. If this upgrade means that the L-population upon reaching the size in question becomes 'infinitely' more preferred in P than L′-populations, this would also account for P ranking this L-population higher than any L′-population, however large. But in the latter case we would still have to assume, I think, that a *further* increase in the size of the L-population would have a lower marginal impact. Which would again imply that, on P, the marginal impact of added lives (L-lives, in this case) will at some point diminish.

assumes, then the Simple View implies that no population of drab lives, however large, could be better than even a *single* excellent life (Jensen 2018).[29] It is an implication that is difficult to accept unless one adheres to a rather extreme form of perfectionism in population axiology. To be sure, Parfit did adhere to a perfectionist view, but it is doubtful that he would have been willing to commit to such a radical version of this position. Therefore, Jensen´s advice to Parfitians is that they should give up the Simple View.

Thus, the following seems to be a fair assessment of Parfit´s proposal: His appeal to incommensurabilities is a suggestion worth serious consideration. Arguably, any plausible attempt to block the Continuum Argument might need to bring in incommensurabilities at some point. As we have seen (cf. the Trilemma above), these incommensurabilities will have to be strictly persistent and thus highly atypical. Still, their persistency can be accounted for if we rely on the modeling of value relations in which incommensurabilities are analyzed in terms of divergent permissible preferences. But if in some of the permissible preference orderings we then need to postulate a life type can be preferentially weakly superior to another life type that is only slightly worse, then we have to give up Parfit´s Simple View. We can no longer insist that the marginal value of added lives never diminishes.[30] This, however, we have reason to do anyway. As Jensen (2018) has shown, we need to discard the Simple View if we want to avoid excessive forms of perfectionism in population axiology. Thus, we must allow for some holistic value effects that can arise when population size increases. We need to allow for diminutions in the marginal value of added lives and perhaps also for its increases in some cases. What sorts of phenomena of this kind it might be plausible to assume remains to be seen.

---

[29] In this proof, Jensen does not directly rely on the Simple View, as that view is stated. Instead he formulates a condition which appears to follow from the Simple View, but which unlike the latter avoids the controversial assumption that values of lives and of populations can be expressed numerically on a cardinal scale – an assumption which Parfit himself otherwise rejects. To formulate this condition, Jensen considers a larger class of possible populations than I have done; the domain of his framework also includes non-homogeneous populations, in which people have lives of different types. His condition can then be formulated as follows:

> For any L and L′, if L is better than L′, then for every k, adding another person with an L-life to kL results in a better population than adding to kL a person with an L′-life.

Intuitively, if the value of adding L-lives were diminishing, then at some point it could be better to add an L′-life to an L-population instead of yet another L-life, despite the fact that L is a better type of life than L′.

[30] This doesn't mean that we should switch to the contrary claim that the marginal value does diminish in a way that makes it possible for a lfe to be weakly superior to a slightly worse type of life. We need not allow for types of life that in *all* permissible preference orderings are preferentially weakly superior to slightly worse types. As we have seen in the preceding section, what is required for strictly persistent incommensurability between L and L′ is that the former is preferentially weakly superior to the latter in some permissible preferences orderings but preferentially *exchangeable* for the latter in other such orderings.

We need to know more about the range of permissible preference orderings of the domain of populations. Thus, there is still work that remains to be done.[31]

References

Andreou, U., 2015, "Parity, Comparability, and Choice", *The Journal of Philosophy* 112: 5–22.

Arrhenius, G., & Rabinowicz, W., 2005, "Millian Superiorities", *Utilitas* 17:127–46.

Arrhenius, G., & Rabinowicz, W., 2015, "Value Superiority", in I. Hirose and J. Olson (eds.), *The Oxford Handbook of Value Theory*, Oxford: Oxford University Press, 225–248.

Griffin, J., 1986, *Well-Being: Its Meaning, Measurement and Moral Importance*, Oxford: Clarendon Press.

Handfield, T., 2014, "Rational Choice and the Transitivity of Betterness", *Philosophy and Phenomenological Research* 89: 584–604.

Handfield, T. and Rabinowicz, W., 2017, "Incommensurability and Vagueness in Spectrum Arguments: Options for saving transitivity of betterness", *Philosophical Studies*, published on-line, August 2017.

Jensen, K. K., 2008, "Millian Superiorities and the Repugnant Conclusion", *Utilitas* 20: 279–300.

Jensen, K. K., 2018, "Weak Superiority, Imprecise Equality and the Repugnant Conclusion", draft.

Parfit, D., 1984, *Reasons and Persons*, Oxford: Oxford University Press.

Parfit, D., 2016, "Can We Avoid the Repugnant Conclusion?", *Theoria* 82: 110–27.

Parfit, D., 2017, "Future People, the Non-Identity Problem, and Person-Affecting Principles", *Philosophy and Public Affairs* 45: 119–56.

[31] Earlier versions of this paper were presented in 2018 at a conference in the memory of Derek Parfit, held in Oxford in May, at a conference on climate change and population axiology held in Stockholm in September, at the School of Philosophy of the Australian National University in Canberra in November and at my own philosophy department at Lund university in December. I am much indebted to the audiences at these events for many useful suggestions. As the Parfit conference provided me with the impetus to write this paper, I'd like to thank its organizers, Joseph Carlsmith, Jeff McMahan and Ketan Ramakrishnan. I am grateful to John Broome, Alan Hájek, Toby Handfield and Karsten Klint Jensen, who have all sent me challenging but very helpful written comments. Apart from making several substantive points, Alan has taken the trouble of hunting down my numerous typos and linguistic inadvertencies. Toby Handfield deserves special thanks; the worry about persistency of incommensurability, which I address and try to dispel in this paper, was first raised in my earlier work with Toby.

Qizilbash, M. (2018), "On Parity and the Intuition of Neutrality", *Economics of Philosophy* 34: 87-108.

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

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

Rabinowicz, W., 2009b, "Incommensurability and Vagueness", *Proceedings of the Aristotelian Society*, Supplementary Volume 83: 71 – 94.

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

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

Temkin, L. S., 2012, *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*, Oxford: Oxford University Press.

Appendix

Trilemma: For any two types of life L and L′ exactly one of the following three relations must obtain: (i) L is exchangeable for L′, (ii) L is weakly superior to L′, or (iii) L admits strictly persistent incommensurability with L′.

Proof:

We want to prove that if L is neither exchangeable for L′ nor weakly superior to L′, then it admits strictly persistent incommensurability with L′.

(This will establish that these three horns of the Trilemma are jointly exhaustive. That they are mutually exclusive is trivial.).

If L is not exchangeable for L′, here must be some quantity k of L such that

(1) No quantity of L′ is better than kL.

But if L is not weakly superior to L′, then there is some quantity k′ of L′ such that

(2) kL is not better than k′L′.

There are three possible cases to that fall under (2):

(i) k′L′ is better than kL.

But this contradicts (1).

(ii) k′L′ is equally as good as kL.

Let k′′ be any quantity larger than k′. By the Principle of Quantity, k′′L′ is better than k′L′. But then, by the transitivity of betterness across equal goodness, (ii) implies that k′′L′ is better than kL. This, however, again contradicts (1).

(iii) k′L′ is incommensurable with kL.

Only (iii) is compatible with both (1) and (2).

So, k′L′ is incommensurable with kL. What about quantities of L′ that are larger than k′? How do they relate to kL?

Given (1), for any quantity k′′ larger than k′, k′′L′ is not better than kL. Nor can it be equally as good as kL (cf. the reasoning above for case (ii)). Can kL be better than k′′L′?

Surely not, given the transitivity of betterness, since by the Principle of Quantity, k′′L′ is better than k′L′, and kL is incommensurable with k′L′.

So, kL is incommensurable not just with k′L′ but also with every quantity k′′ of L′ larger than k′.

This means that

> (3) kL is incommensurable with every sufficiently large quantity of L′ (and better than all smaller quantities).

The same conclusion can be established for any quantity k^+ of L larger than k. By the Principle of Quantity and the transitivity of betterness,

> (1^+) No quantity of L′ is better than k^+L.

For then this quantity of L′ would also be better than kL, which contradicts (1).

And, since L is not weakly superior to L′, there must be some quantity k′ of L′ such that

> (2+) k^+L is not better than k′L′.

But then, by the same argument as above, k^+L must be incommensurable with k′L′ and with every quantity of L′ that is larger than k′.

Thus,

> (4) For all k^+ larger than k , k^+L is incommensurable with every sufficiently large quantity of L (and better than all smaller quantities).

(3) and (4) imply that

> (5) L admits strictly persistent incommensurability with L′.

Q. E. D.

Gustaf Arrhenius & Julia Mosquera[1]

Positive Egalitarianism

According to Positive Egalitarianism, not only do relations of inequality have negative value, as Negative Egalitarians claim, but relations of equality also have positive value. The egalitarian value of a population is a function of both pairwise relations of inequality (negative) and pairwise relations of equality (positive). Positive and Negative Egalitarianism diverge, especially in different number cases. Hence, an investigation of Positive Egalitarianism might shed new light on the vexed topic of population ethics and our duties to future generations. We shall here, in light of some recent criticism, further develop the idea of giving positive value to equal relations.

[1] Institute for Futures Studies; Gustaf.Arrhenius@iffs.se, Julia.Mosquera@iffs.se. Gustaf Arrhenius and Julia Mosquera contributed equally to this work.

I. Introduction

The topic under discussion here, following in the tracks of Larry Temkin's influential work on inequality, is how to rank populations with regard to their egalitarian value, that is, how they can be ordered by the relation "is at least as good as with respect to egalitarian concerns".[2] One should distinguish this undertaking from the project of how such a ranking would play into the all things considered rankings of populations, where we also have to consider other aspects, such as the total welfare in a population, and from the project of ranking populations in terms of the primarily descriptive relation "is at least as (un)equal as".[3] Another way of putting it is that we are asking what our well-informed preference from a moral perspective would be if we only cared about inequality and equality (of some sorts).[4]

Several authors, for example Larry Temkin, Ingmar Persson, and Shlomi Segall, have argued in favour of the view that inequality among people is bad but equality among people is neither good nor bad but of neutral value.[5] It is fair to say that this is the received view on the value of inequality and equality. (Arrhenius 2013) formulated a version of this view with a bit more content:[6]

> *Negative Egalitarianism*: The egalitarian value of a population is a strictly decreasing function of pairwise relations of inequality.

Negative Egalitarianism can be contrasted with a novel view that, in addition to the negative value of unequal relations, also ascribes positive value to relations of equality, as suggested by (Kawchuk 1996), (Arrhenius 2013), and (Mosquera 2017b). We shall refer to this position as Positive Egalitarianism:

---

[2] Cf. (Temkin 1993; Arrhenius 2009, 2013). For simplicity, we are here only considering human populations but the framework could easily be extended to include other sentient beings.

[3] We say "primarily" since this relation might be partly evaluative because the currency might involve an evaluation. For example, to say that Tim has higher welfare than Orri in outcome X is an evaluation to the effect that Tim is *better off* than Orri in X. For the same point, see (Rabinowicz 2003), fn. 7.

[4] See (Arrhenius 2013) for a more detailed discussion of this topic.

[5] (Persson 2001), p. 30, holds the same view: "According to this conception [of egalitarianism], the intrinsic value of just equality will be neutral, consisting in the mere absence of something intrinsically bad, namely unjust inequality". Larry Temkin's position seems to be similar. Although Temkin doesn't explicitly address the question of whether equality has positive value, he provides a number of arguments for why the mere addition of equally well-off people doesn't improve a population from the point of view of equality, but rather worsens it, since such addition can increase the inequality in a population in various different respects. See for example (Temkin 1993), pp. 23, 211. See also (Segall 2016), pp. 74-86.

[6] See (Arrhenius 2013). Negative Egalitarianism wasn't separately stated in Arrhenius's paper but it makes up the "inequality part" of his definition of Positive Egalitarianism (p. 85).

> *Positive Egalitarianism*: The egalitarian value of a population is a strictly decreasing function of pairwise relations of inequality and a strictly increasing function of pairwise relations of equality.[7]

Let's us expand on what these two positions involve. First, they presuppose a currency of egalitarian justice which is a contested subject.[8] For simplicity, let's assume that the egalitarian currency is welfare (or, to put it differently, "welfare" is a place-holder for whatever is the correct currency of egalitarian justice) and that individual welfare is measured on an interpersonally comparable ratio scale. Moreover, again for simplicity, let the measure of pairwise relations of inequality be the absolute value of the difference in welfare between the two involved individuals.[9]

Consider a population A consisting of three persons, p_1, p_2, p_3, with welfare 10, 20, and 20 respectively. We can represent this as A = <10, 20, 20>. According to Negative Egalitarianism, the egalitarian value of A is a function of the absolute value of the welfare differences of all distinct pairs of different individuals in A, that is, (p_1, p_2), (p_1, p_3), and (p_2, p_3).[10] In this case, there are two pairwise relations of inequality of size 10 between p_1 and p_2, and p_1 and p_3 respectively, which are of negative value. According to Positive Egalitarianism, the egalitarian value of A isn't only a function of the two inequalities of size 10 but also of the equality between p_2 and p_3, which is of positive value.

Negative and Positive Egalitarianism can diverge in both same and different number cases. An example of a same number case is that Negative Egalitarianism will always consider populations like A bad from the perspective of egalitarian concerns, whereas some version of Positive Egalitarianism might evaluate such populations as good, e.g., when the involved inequality is so small that its negative value is outweighed by the positive value of the equal relation. As we shall see, however, these views come apart even more in different number cases. Moreover, these views are substantial positions that rule out some positions regarding the value of equality (more on this below).

There are at least three good reasons for why considering egalitarian aspects of

---

[7] (Arrhenius 2013).

[8] For this debate, see e.g., (Rawls 1971), (Sen 1992b, 1993, 1980), (Dworkin 1981a, 1981b, 2000), (Cohen 1989, 1993), (Arneson 1989), and (Nielsen 1996).

[9] One might also reasonably make this measure more sophisticated to take into account that, for instance, differences at low levels matter more than differences at high levels (by, for example, transforming the individual welfare levels with a strictly increasing concave function). There are other more sophisticated functions that could be used but since this won't be relevant for the discussion below, we shall leave this issue aside here.

[10] Hence, it's not a function of the identity pairs, e.g., (p_1, p_1), and each pair of distinctive individuals are only counted once, e.g., (p_1, p_2) and (p_2, p_1) is counted as one pair.

population change and different number cases is interesting. Firstly, doing so might shed new light on the vexed topic of population ethics and our duties to future generations since equality is one aspect in the ranking of future populations of different sizes. Secondly, if we want to measure the value of equality and inequality over time, we have to take into account that the size of the compared populations might have changed. For example, when one tries to measure the development of global equality and inequality during the last thirty years or so, one has to take into account the great population expansion in countries such as India and China.[11] Moreover, climate change is very likely to affect not only future peoples' welfare but also the size of future populations.[12] Thirdly, this new dimension of egalitarian theory is a fruitful way of probing our ideas about egalitarian concerns and reveals as yet underdiscussed and surprising complexities and problems in our current conceptualization of the value of equality, as the discussion above indicates.

II. Mere Addition of Better off People

Let us fist consider a different-number case from (Arrhenius 2013):

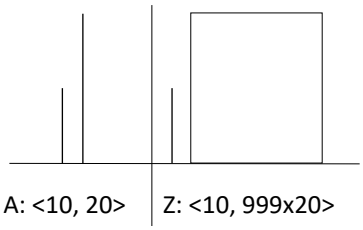


Diagram 1. Mere Additions of Better off People

In Diagram 1 above, the height of the vertical lines and the rectangle represent people's welfare according to some suitable measure or proxy. Each vertical line represents one individual whereas the rectangle represents a group, the size of which is indicated by the width of the rectangle. All the lives in the above diagram have positive welfare, or, as we also could put it, have lives worth living, which is represented by positive numbers.[13] Population A consists of two persons, one with

[11] See, e.g., (Bosmans, Decancq, and Decoster 2014) for some important results in this area.
[12] For a discussion of some of the relationship between climate change and demography, see (Hales et al. 2014; Bommier, Lanz, and Zuber 2015; Mejean et al. 2017; T. Carleton et al 2018; Geruso and Spears 2018).
[13] For a discussion of alternative definitions of a neutral life, many of which would also work fine in the present context, see (Arrhenius 2000b, forthcoming, ch. 2 and 9). See also (Broome 1999) (Broome

half of the welfare of the other, 10 and 20 units respectively. In Z, we have added another 998 best off people with 20 units of welfare.

Intuitively, Z seems to be better with respect to egalitarian concerns since Z is very close to perfect equality. A, on the other hand, is a straightforwardly unequal population. Here's how we can analyse the above intuition. In A, there are no relations of equality but one relation of inequality. In Z, on the other hand, there is a staggering 498 501 relations of equality, which in comparison outweighs the puny 999 relations of inequality.[14] Hence, one might reasonably judge Z as better than A in regard to egalitarian concerns, and a natural way to account for this intuition is to claim that from an egalitarian perspective, we should not only care about relations of inequality but also about relations of equality, as suggested by Positive Egalitarianism.

Notice that it doesn't follow from Positive Egalitarianism that Z is better than A in respect to egalitarian concerns since it is compatible with any ranking of these populations. The ranking depends on what weight we give to the great increase in equal relations as compared to the relatively small increase in unequal relations.

Interestingly, that Z is better than A in regard to egalitarian concern is implied by the Gini Coefficient (and some other standard economic measures, for example, Relative Mean Deviation).[15] The Gini Coefficient is approximately 0.17 for A and only 0.0005 for Z.[16]

Negative Egalitarianism, on the other hand, implies that Z is worse than A with regard to inequality since in the change from A to Z, the number of unequal relations increases, from 1 to 999, and the size of the gaps are the same.[17] This shows, interes-

---

2004), and (Parfit 1984, 357–58), and appendix G).

[14] Since the number of pairwise relations in a population of size n is $\frac{1}{2}n(n\text{-}1)$, we get that the number of equal relations in Z is $999(999\text{-}1)/2 = 498501$.

[15] See (Kawchuk 1996,p.159), (Arrhenius 2013, 2016) for a discussion of such measures. One might also be interested in *Average Per Pair Inequality (APPI)*: The sum of the absolute value of all welfare differences for all distinct pairs of individuals in the population divided by the number of such pairs (see (Rabinowicz 2003) and (Arrhenius 2013)). APPI also ranks Z as better than A with respect to inequality since $APPI(A) = (|10\text{-}20|)/2(2\text{-}1)/2 = 10$ whereas $APPI(Z) = (999|10\text{-}20|)/(1000*999)/2 = 0.02$. A measure proposed by Derek Parfit is likely to have the same implication. Parfit compares two populations, A+ and Alpha. A+ consists of two groups of people of the same size, one with 100 units of welfare per person, and one with 50 units of welfare per person. Alpha consists of one group of the same size as A+ but with 105 units of welfare per person and a very large group of people with 45 units of welfare per person. He writes: "The inequality in Alpha is in one way worse than the inequality in A+, since the gap between the better-off and the worse-off people is slightly greater. But in another way the inequality is less bad. This is a matter of the relative numbers of, or the *ratio* between, those who are better-off and those who are worse-off. Half of the people in A+ are better off than the other half. This is a worse inequality than a situation in which almost everyone is equally well off, and those who are better off are only a fraction of one per cent. - - - All things considered, the natural inequality in Alpha is not worse than the natural inequality in A+." (Parfit 1986, p.156).

[16] The Gini Coefficient is defined as: $\frac{1}{2n^2\hat{y}}\sum_{i=1}^{n}\sum_{j=1}^{n}|y_i - y_j|$.

[17] Likewise according to Temkin's Individual Complaint Theory (Temkin 1993, ch. 2) since the worst off person has many more complaints in Z as compared to in A, 999 complaints versus 1.

tingly, that the Gini Coefficient is ruled out by Negative Egalitarianism. The reason is that Gini not only takes into account the number of unequal pairs but is also sensitive to the number of equal pairs in a population since they figure in the denominator. Hence, Gini isn't, as one might believe, a measure that only gives weight to unequal relations, but it rather belongs to the family of positive egalitarian theories that also give weight to equal relations.[18]

(Segall 2016) has considered this case and disagrees with the judgement explained above of positive egalitarians and others who rank Z as better than A in regard to egalitarian concerns:

> In A there are two individuals living in inequality. But in Z the sense of inequality of that one worse off person is magnified (almost) a thousand times. And the fact that there are 999 better-off individuals living in equality amongst themselves does not detract from that badness, and potentially even emphasises her isolation as a worse-off individual in that society.[19]

However, Positive Egalitarianism doesn't claim that the "999 better-off individuals living in equality amongst themselves ... detract from ... [the] badness" of the inequality. On the contrary, Positive Egalitarianism implies that Z is worse in one respect that is relevant for egalitarian concerns, namely in respect to the number of unequal relations to which the worst off person is subjected. It is just that positive egalitarians also value equal relations, and since there are so many of them in Z, this may outweigh the badness of the unequal relations when it comes to egalitarian concerns. Again, in A, there are no relations of equality and one relation of inequality, whereas in Z there is a staggering 498 501 relations of equality, which in comparison outweighs the few 999 relations of inequality.

Secondly, let us again stress that it doesn't follow from Positive Egalitarianism that Z is better than A in respect to egalitarian concerns, but it is compatible with that judgement. It depends on what weight we give to the increase in unequal relations as compared to the increase in equal relations. Hence, Positive Egalitarians could accommodate Segall's intuition by changing the relative weight of the value of equal and unequal relations. To make Segall's claim into a real argument against Positive Egalitarianism, one would have to show that there are no cases of

[18] Strictly speaking, Gini is ruled out by Positive Egalitarianism as we have defined it above since adding equal relations to a perfectly equal population won't increase the egalitarian value according to Gini (it will be the same) whereas this follows from Positive Egalitarianism. Gini is, however, compatible with a weaker version of Positive Egalitarianism according to which the egalitarian value of a population is a decreasing function of pairwise relations of inequality and an increasing function of pairwise relations of equality.

[19] (Segall 2016), p. 76

the type depicted in Diagram 4, that is, cases in which the positive value of an increase in equal relations outweighs the negative value of a much smaller increase in unequal relations, possibly involving very small inequalities. This is a rather tall order, we surmise, since it is tantamount to showing that irrespective of how small the inequalities are and how many more equal relations there are, the positive value of the equal relations cannot outweigh the negative value of the unequal relations.

Segall's point regarding the isolation of the worst off can be understood in different ways. It could be taken as a psychological claim regarding the welfare of the worst off individual, that she suffers from being the only worst off person. Taken this way, the objection misses its target since any effects on the welfare of the person is already taken into account in the diagram and the worst off has the same welfare in both outcomes.

More interestingly, Segall hints at the putative badness of the isolation of a single worse off individual in a society in the quote above. This might be a further aspect to take into account when we rank A and Z with respect to egalitarian concerns. (Mosquera 2017a) has a more developed proposal that also takes account of this intuition. Her idea is that the way in which relations of inequality are distributed should also matter when accounting for the egalitarian value of a population. It is for example pro tanto worse if the inequalities are concentrated on a single individual rather than on many individuals. Mosquera refers to this as "Distribution-Sensitive Pairwise Egalitarianism". Going back to the case with A and Z, there are more relations of inequality concentrated on the worse off individual in Z, as compared to A. In this respect, Z is worse than A.

However, the question is what weight should be assigned to those considerations as compared to other egalitarian considerations. Mosquera doesn't commit to any particular weighing and would not claim that her account implies that Z is worse than A, taking into account all the relevant egalitarian concerns. It is indeed possible for an egalitarian to assign such an extremely high weight to the putative badness of the 'isolation' of a single worse-off individual in a society, such that the ranking of A and Z in respect to egalitarian concerns would be reversed as compared to the ranking provided by Positive Egalitarianism and Gini. But it is hard to find a convincing argument for such a radical view, and very few people would be attracted to such a view, we surmise.

To sum up, we think that Segall's ranking of A and Z is partly based on an equivocation of inequality and egalitarian concerns. Furthermore, it overlooks the great number of the equal relations in Z as compared to the small number of unequal relations. It is this difference that makes it intuitively attractive to rank Z as better than A in respect to egalitarian concerns and along the lines of reasonable versions of Positive Egalitarianism and other theories such as the Gini Coefficient. The

appeal to 'isolation' or 'concentration' isn't sufficient to reverse the ranking in this case.

III. The Progression Case

As we showed above, Negative and Positive Egalitarianism come apart in different-number cases. But as we also mentioned in the introduction, Negative and Positive Egalitarianism can also diverge in same-number cases. In fact, Segall suggests another criticism against Positive Egalitarianism that is based on a same-number case where he claims that Positive Egalitarianism yields the wrong result. It might be instructive to look at it now. Consider the following two same-sized populations of a thousand people each:

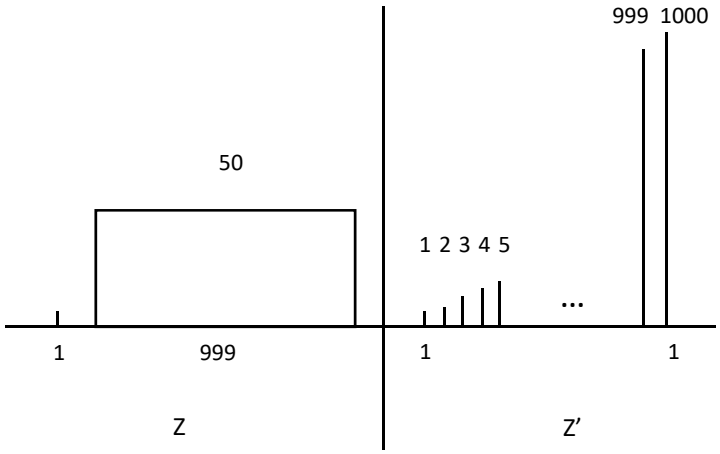


Diagram 2. The Progression Case

In diagram 2, the numbers above the height of the vertical lines and the rectangle represent people's welfare. The number of people at each welfare level is indicated by the numbers below the neutral welfare line.

Segall claims that Z' is better than Z in regard to egalitarian concerns:

> My intuition is that Z' is better than Z with respect to equality. This is so even though Z' contains many more incidents of inequality, and no relations of equality at all.[20]

[20] (Segall 2016), p. 76

Positive Egalitarianism of course yields the opposite ranking. One might get the impression from Segall's discussion that what makes Z better than Z' according to Positive Egalitarianism is the fact that Z contains a lot of relations of equality, while Z' has no relations of equality. As we shall show below this is wrong since Negative Egalitarianism also ranks Z as better than Z' in respect to egalitarian considerations. Positive Egalitarianism gives extra reasons for this ranking because of the positive value of the relations of equality in Z, but it is not the main reason for why Z is better than Z'.

It might not be evident that both Negative and Positive Egalitarianism imply that Z' is worse than Z with respect to equality so let us informally demonstrate it. We can do a one-to-one mapping of all the 999 unequal pairs in Z to a corresponding unequal pair in Z' with a greater gap. According to Negative Egalitarianism, this subset of unequal pairs from Z' will be worse with respect to equality than the set of unequal pairs from Z. Since all the pairs in Z' that remain unmapped after this are unequal, this will only further decrease the egalitarian value of Z'. The same demonstration holds for Positive Egalitarianism since the equal relations in Z will increase the egalitarian value of Z further, as compared to the egalitarian value of Z'. Here's is a figure illustrating the same point:

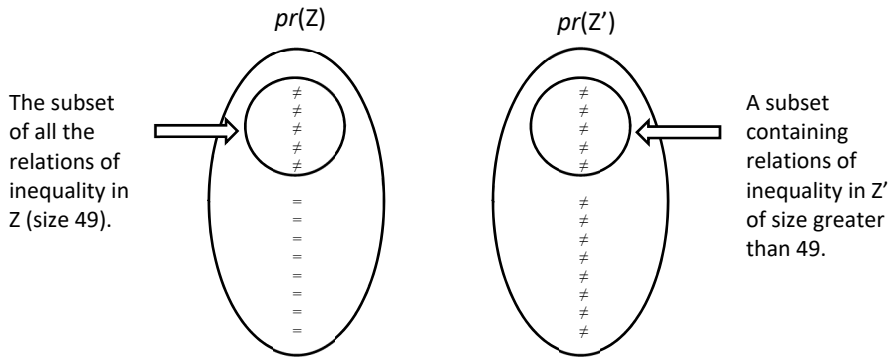


Diagram 3. A one-to-one mapping of the unequal relations in Z to unequal relations of greater size in Z'

In the diagram above, the two ellipses denoted *pr*(Z) and *pr*(Z') represent the set of all pairwise relations of equality and inequality in populations Z and Z' respectively. Relations of equality are represented by "=" whereas relations of inequality are represented by "≠". The horizontal arrows from *pr*(Z) to *pr*(Z') represent the mapping of all the relations of inequality from *pr*(Z) to relations of inequality of

greater size in *pr*(Z'). Hence, the circle in *pr*(Z) represents the set of all relations of inequality in *pr*(Z) whereas the circle in *pr*(Z') represents some subset containing relations of inequality in *pr*(Z') of size greater than 49. The latter subset will be worse with respect to equality than the former according to Negative Egalitarianism. As we can see in the diagram, all the pairs in *pr*(Z') that remain unmapped are unequal, and thus will further decrease the egalitarian value of Z'. Moreover, all the unmapped pairs in *pr*(Z) are equal and thus will increase the egalitarian value of Z further as compared to the egalitarian value of Z'.

Does Segall's case show that we have to reject both Positive and Negative Egalitarianism? We don't think so. Segall's ranking of Z and Z' overlooks the great magnitude of the inequalities in Z'.

Firstly, as Segall concedes, population Z' contains a greater number of relations of inequality than Z.[21] However, it is important to notice how much greater the number of such relations is in Z' as compared to Z. As Mosquera (2017) has pointed out, the latter population contains only 999 relations of inequality of size 49 whereas Z' contains a whopping 499 500 relations of inequality of sizes ranging from 1 to 999.[22]

Secondly, the vast majority of the relations of inequality in Z' involve a bigger gap than the inequalities in Z. In Z', 450 775 relations out of a total of 499 500 relations of inequality are of size greater than 50. In Z, on the other hand, the relatively puny 999 relations of inequality are all of size smaller than 50.[23]

As an illustration only of the magnitude of the inequality contained in Z', consider individual 1 in Z' who is subject to 949 relations of inequality greater than 49. As pointed out by Mosquera (2017), these 949 relations of inequality range from size 50 to size 999, whereas, again, the relations of inequality in Z are of size 49.[24]

Hence, there are many more and greater inequalities in Z' as compared to Z. Thus, pace Segall, it makes eminent sense for an egalitarian to rank Z' as worse than Z in respect to egalitarian concerns.

Finally, let us here take the opportunity to point out a simple misunderstandding of Positive Egalitarianism on Segall's behalf. He writes:

---

[21] The formula for calculating the number of pairwise relations in a population of size N is N(N-1)/2. In the case of Z', where there are no equal relations, the number of unequal relations coincide with the total number of pairwise relations in the population as defined above. In the case of a population with both unequal and equal relations, the number of unequal relations equal the total number of relations minus the number of equal relations.

[22] See (Mosquera 2017b, 186).

[23] Moreover, of the 1000 individuals in Z', 949 individuals are subject to at least one relation of inequality greater than 50 (which represents 94.9% of population Z') as opposed to 1 individual in Z (which represents only 0.1% of the population).

[24] See (Mosquera 2017b, 186).

> A major difference [between Segall's and Positive Egalitarianism's verdicts on Z and Z'] of course is that Arrhenius's positive egalitarianism *simply counts the number of incidents of both equality and inequality but does not take into consideration the magnitude of the latter*.[25]

Positive Egalitarianism, however, does take into account the size of the inequalities in the distribution since it is an extension of Negative Egalitarianism. The set of all pairs of individual welfare levels in a population contains information about the size of the gaps and the number of unequal pairs. Negative and Positive Egalitarianism take into account both of these aspects so that the greater the gap, and the more of such unequal relations, the worse in regard to egalitarian concerns. This is rather clear from Gustaf Arrhenius' original paper in which there is extensive discussion on how to take into account the size of inequalities, but also on how to weigh them against the positive value of equal relations.[26] For example, Arrhenius writes that "it seems appropriate that both the number and the size of inequalities matter".[27]

More importantly, Negative and Positive Egalitarianism, as defined above, rule out egalitarian views according to which the egalitarian value of a population sometimes isn't affected by an increase in the gaps or by an increase in the number of unequal pairs. A simple example is the range measure where the egalitarian value only depends on the gap between the best and worst off in the population. On this view, the egalitarian value doesn't change with an increase in the number of pairwise inequalities as long as the distance between the best and worst off is preserved (by, for example, adding more worse off people). Likewise, the view that Segall ascribes to Arrhenius above, which ignores the size of the inequalities and only counts the number of unequal pairs,[28] is actually incompatible with Negative and Positive Egalitarianism.[29]

---

[25] (Segall 2016), p. 76., last emphasis added.

[26] For just one example of the latter, see Arrhenius (2013), pp. 85-86, where he discusses how to weigh the goodness of an increase in equal relations against an increase in in slightly unequal relations. There are many more examples.

[27] (Arrhenius 2013), p.79.

[28] Mosquera labels this kind of view "Aggregative Pairwise Inequality" in (Mosquera 2017b, 149).

[29] This feature makes Positive and Negative Egalitarianism unsuitable as a general characterisation of all egalitarian views and shows that they're substantial positions. For the former we would need something weaker not to exclude positions such as the range measure. Indeed, Negative Egalitarianism rules out Gini in different number cases since, as shown by (Arrhenius, mimeo), according to Gini, a mere increase in unequal relations by adding more individuals to the population might sometimes improve the situation in respect to inequality. Another principle that Negative Egalitarianism rules out is Temkin's BOP (The Relative to the Best-Off Person View of Complaints) in combination with his Additive Principle of Equality (Temkin 1993). According to this combined principle, all the individuals but the very best off have a complaint and the badness of the inequality of a population is equivalent to the aggregated number of complaints of all but the very best off in that population. A principle like this, in difference to Negative Egalitarianism, is insensitive to the addition of better off people (i.e., adding more people at the level of the better off doesn't increase the badness of inequality) since it only looks at

IV. Class Division

Are there any other reasons to rank Z as worse that Z' with respect to egalitarian concerns? (Segall 2016) again hints at the putative badness of the isolation of a single worse-off individual in a society (see quote in the preceding section). But analogously to our reasoning above regarding the case in Diagram 1, this putative badness can hardly outweigh the greater number and the greater size of the inequalities in Z' as compared to Z.[30]

More interestingly, one might argue that Z is worse than Z' regarding another egalitarian consideration, namely class division. In Z, there is a clear class division with one group of better off people and another group with (only) one worst off person. In Z', one might argue, there is no clear class division since there is only a small difference in wellbeing between each adjacent individual (see Diagram 2). Hence, there seems to be no distinct upper or lower classes in Z', or so the argument goes.[31]

Several things can be said in response to this argument. Firstly, whether Z and Z' can be interpreted as containing a class division depends on the conception of welfare involved in the case or whether an index or proxy for welfare is used. For some conceptions or proxies of welfare, welfare differences might very well track class differences, understood in terms of social or economic status, the most common elements used to define class.[32] For others, there is no clear connection. For example, conceptions or proxies of welfare such as income and wealth are likely to track class differences.

On the other hand, with conceptions of welfare such as experientialist, desire satisfaction, and objective lists theories the connection between welfare and class is weaker. For example, given these conceptions, the reasons for why the welfare of individual 1 in Z is significantly lower than the others' may be due to a painful medical condition that others in that population don't suffer, but she might otherwise enjoy high social and economic status. She might, for example, be a mem-

---

the difference with the best-off person, irrespective of the number of people who happen to be at this level. Although a general characterisation of all egalitarian views is an interesting question, we set this issue aside for now.

[30] Moreover, one might argue that there is isolation in population Z' too since no one has the same level of welfare. On the other hand, the badness of isolation might depend on how big the welfare difference is. We shall not pursue this line of argument further here.

[31] As suggested by Shlomi Segall in personal communication.

[32] Notice that there isn't consensus on what is the right definition of 'class'. While contemporary political scientists tend to identify class with income or wealth (i.e., they refer to class differences mainly as income or wealth differences), sociologists tend to include other elements in the definition, such as educational background, parental background, and/or the relation to the means of production. We would like to thank Malcolm Fairbrother, Chandra Kumar, and Stefan Svallfors for a discussion of this issue.

ber of a wealthy aristocratic family despite suffering from such a painful condition. Such interpretations are not ruled out by the very abstract description of the case that we have provided here. In conclusion, these conceptions of welfare differences don't always track class differences.

The notion of welfare used in this paper is a broad one and is compatible with experientialist, desire satisfaction, and objective list theories.[33] Given that we use a broad notion of welfare, one cannot infer from the description of Z and Z' that those populations contain any particular social stratification, or that the individuals involved in those populations belong to any particular class.[34]

However, let's for the sake of the argument accept that Z and Z' can be compared in terms of class division. Even so, it would not be true that Z is worse than Z' from the point of view of class division. This becomes clearer with the following representation of the case:

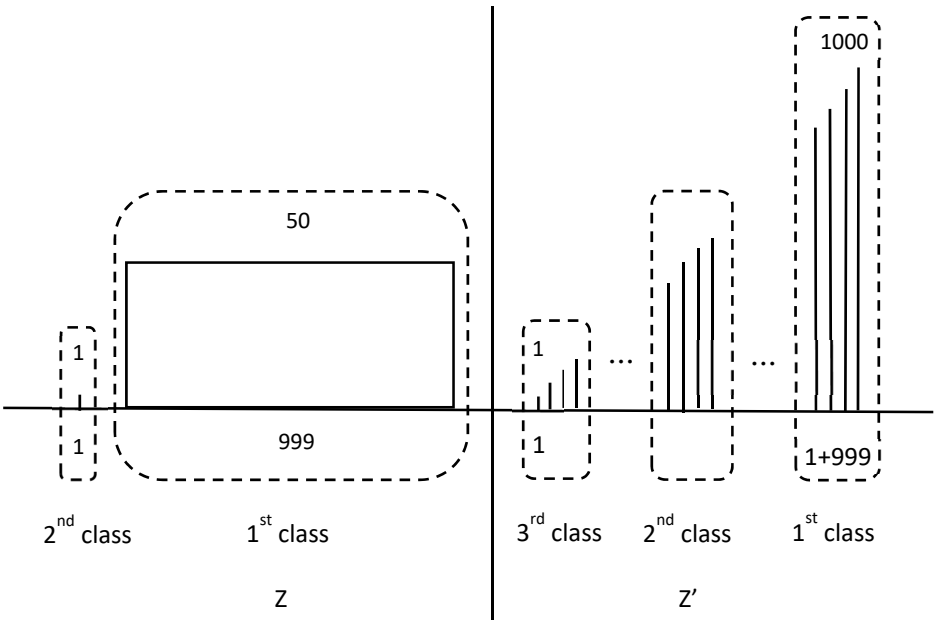


Diagram 4. Possible class divisions in Z and Z'

[33] For experientialist theories, see e.g., (Sumner 1996), (Feldman 1997, 2004), and (Tännsjö 1998). For desire theories, see e.g., (Barry 1989), (Bykvist 1998), (Griffin 1990), and (Hare 1981). For objective list theories, see e.g., (Braybrooke 1987), (Hurka 1993), (Rawls 1971), and (Sen 1980, 1992a, 1993).
[34] One might also object to the conceptual possibility that a single person can constitute a 'class' (the single person at low level in Z). We shall not pursue this line of argument further here.

Although the difference in welfare between each pair of adjacent people in Z' is small, by looking at diagram 3 above we can clearly see that there are groups whose individuals are much better off than the individuals in other groups, such as the people in the 1st and 2nd class as compared to the people in the 3rd class. Indeed, we can imagine that these groups represent the lower class, middle class, and the upper class in that population. And the fact that there are people in between these three classes that cannot be clearly classified as belonging to any of these classes doesn't refute the existence of this class division in Z'.

Perhaps one might object here that it is not clear how the grouping into classes would be made in a society that looked like Z', or that there is no salient way of doing it, while such a grouping is rather natural in Z. As true as this might be, it is of no normative significance. That there are many ways of grouping people doesn't show that there wouldn't be some ways (or at least one) that are more relevant from the perspective of egalitarian concerns. And our point is that there is at least one such grouping that shows that Z' is also a class society. Hence, it is not the case that Z is worse than Z' regarding this egalitarian consideration.

V. Perfect and rough equality

At this point, one might worry that the question of whether equality can have positive value is only interesting if people can in fact have *exactly* the same level of the currency that ought to be distributed equally, in a wide range of cases. For example, in realistic cases, perfect equality of welfare will almost never hold, or so one might argue. Now, if welfare is the correct egalitarian currency, and if it is so rare that two or more individuals will be perfectly equal in respect to welfare, what is then the point of determining the positive value of equality? Doesn't this make the focus on equality albeit theoretically interesting, practically irrelevant? Larry Temkin has recently put forward this argument.[35]

There are at least two answers to Temkin's worry. Firstly, whether people can achieve perfect equality depends on how the egalitarian currency is specified. In realistic cases, perfect legal equality (e.g., the same laws for different ethnic groups) is arguably much easier to obtain than perfect equality of welfare and has already been achieved among large groups of people in many countries, or so one might argue. Hence, if one cares about legal equality, then the value of equality is practically relevant. Moreover, one could care about inequality and equality for several currencies at the same time, for example both legal equality and equality of welfare.

[35] In personal communication.

In that case, even if perfect equality of welfare is rare, the value of equality is still practically relevant since perfect legal equality isn't rare.

Secondly, Arrhenius (2013) has suggested a version of Positive Egalitarianism in terms of what he calls "rough equality". Small differences in welfare, that is, small inequalities, can equally well be described as "rough equalities". One might then ask whether it is only perfectly equal relations that have positive value or if this is also true of roughly equal relations. A promising idea is that inequality has negative value when the inequality is sufficiently big ('strict inequality' as we can call it) but when it gets smaller we reach the border for rough equality, which is of neutral value, and when the welfare difference gets even smaller, rough equality has positive value, and this value increases the closer we get to perfect equality. So, for example, <5, 10> might have negative value from an egalitarian perspective; <8, 10> neutral value since we have reached the border for rough equality; <9, 10> positive value since we are getting close to perfect equality; and lastly <10, 10> has maximal positive value regarding egalitarian concerns.

We can formulate Rough Positive Egalitarianism in the following way:

> *Rough Positive Egalitarianism*: The egalitarian value of a population is a strictly decreasing function of pairwise relations of strict inequality and a strictly increasing function of pairwise relations of rough and perfect equality.[36]

This view, we suggest, can accommodate the worry raised by Temkin. Rough Positive Egalitarianism incorporates the idea that individuals don't need to be at the *exact* same level of equality in order for there to be some positive value of equality in the relation that holds between them. There is a point at which individuals' welfare is close enough to each other's to elicit this value. This view, we believe, fits the natural way in which people think about equality. When people claim that people ought to be equal, they rarely mean that people ought to be perfectly equal but rather, roughly equal, we surmise.

Segall is one of the few authors who has discussed Rough Positive Egalitarianism. He has put forward a claim that involves Z and Z' (the populations involved in the Progression Case discussed in the previous sections). According to Segall, Rough Positive Egalitarianism judges Z as worse than Z'. And this, Segall argues, contradicts the ranking of Positive Egalitarianism for these populations:

> There are reasons to think that Arrhenius himself ought to also be committed to the view that Z' is better than Z with respect to equality. When considering the

[36] Cf. (Arrhenius 2013), p. 89.

> badness of small inequalities he says that a promising approach is 'that inequality has a negative value when the inequality is sufficiently big but, when it gets smaller, we have rough equality that is of neutral value'.

And he continues, referring to the supposed implications of Rough Positive Egalitarianism:

> On this reading [of Rough Equality], in Z we have 999 relations of sufficiently big inequalities and hence negative value, whereas in Z' we have almost half a million relations of small inequalities and hence neutral value. So it seems that even on Arrhenius' intuition Z' is better than Z, even though the former contains no relations of equality.[37]

It is important to note, though, that it is not the case that "in Z' we have almost half a million relations of small inequalities and hence neutral value". Sure, there are small inequalities between adjacent or close pairs in the progression and they might be of neutral value. But, as we pointed out above, the vast majority of the relations of inequality in Z' involve big and greater gaps than the inequalities in Z. Hence, we can conclude that also for those who adhere to Rough Positive Egalitarianism, there are very strong reasons to hold that Z' is worse than Z with respect to egalitarian considerations. The same reasons that support the ranking provided by Negative and Positive Egalitarians also apply for Rough Positive Egalitarianism.

VI. The Repugnant Equality Objection

Finally, one may think that upholding to a view like positive egalitarianism, which attributes positive value to relations of equality, might lead to a version of Derek Parfit's famous Repugnant Conclusion, which, following Mosquera (2017b), we will call "Repugnant Equality":

> *Repugnant Equality*: For any population consisting of people enjoying the same very high positive welfare, there is a better population with respect to egalitarian concerns in which everyone has the same very low positive welfare.[38]

[37] (Segall 2016), p. 76, fn. 7.
[38] Our formulation. See (Segall 2016) and (Mosquera 2017b) for a similar formulation.

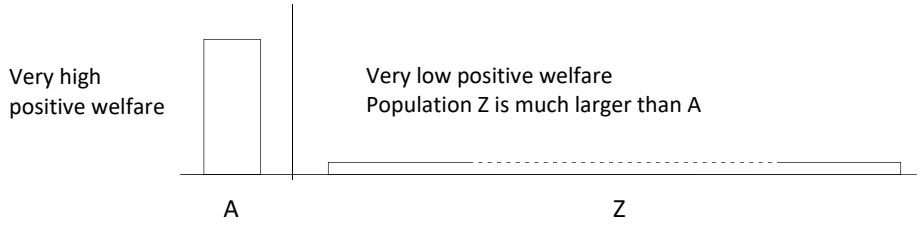


Diagram 5. Repugnant Equality

According to Segall, Positive Egalitarianism "probably" ranks Z as better than A from the point of view of equality even when the lives in Z are barely worth living since Z "contains many more relations of equality" as compared to A.39 Hence, Positive Egalitarianism implies that Z is better than A in one respect, namely in respect to egalitarian concerns, that is, Repugnant Equality. He finds this implication counterintuitive.[40]

Repugnant Equality is a variation of Derek Parfit's well-known Repugnant Conclusion.41 It is also, in a way, a version of the levelling down objection. According to the levelling down objection, one cannot make a situation better *in any respect* by only decreasing welfare of the better off (to the same level as the worse off). Many conceptions of the value of equality, like those discussed here, have this implication, but, along with many egalitarians, we don't find the objection persuasive. It would be a different matter if a theory implied that it was all things considered better to level down, which of course would be highly counterintuitive.[42] Likewise, it doesn't follow from Repugnant Equality that one has to accept that Z is better than A all things considered but only that Z is better in one respect. We reject the first position but, pace Segall, don't find the other position counterintuitive.

The very idea that it can be a decisive argument against a theory that it implies that an outcome is better or worse *in one respect* is often question-begging. The reason is that this kind of objection can be launched against any theory according to which the value can vary in at least two dimensions. This is, however, an unsurprising feature of such theories. Take Hedonistic Total Utilitarianism, according to which the value of an outcome is determined by the total sum of pleasure over pain,

[39] (Segall 2016), p. 77.
[40] (Segall 2016), p. 77.
[41] See (Parfit 1984), p. 388.
[42] Cf. (Arrhenius 2013), p. 92. (Segall 2016), p. 78, agrees with this: "Levelling down would bring about equality. This is, for sure, a defeasible reason, but it's a reason nevertheless." For a further discussion of the levelling down objection where he, correctly in our view, argues that a theory "may judge a scenario to make an outcome better (in one respect), even when it's better for no one, is not self-evidently problematic". (Segall 2016), p. 350

as an example. Consider the value of adding a person with negative lifetime total sum of pleasure and pain. Of course, this would be all things considered bad according to Hedonistic Total Utilitarianism. However, assume that this life involves one minute of pleasure in addition to all the pain and suffering. Then it is true that in one respect, from the perspective of Hedonistic Total Utilitarianism, it is good to add this life. Clearly, this isn't much of an objection to this version of Utilitarianism. Likewise for the levelling down objection against egalitarian theories in general and the 'in-one-respect-better objections' raised above by Segall against Positive Egalitarianism. There is nothing counterintuitive about this feature of multidimensional axiologies for which outcomes can be both good and bad in different respects.

Moreover, positive egalitarians are not even committed to ranking Z as better than A in respect to egalitarian concerns, that is, to Repugnant Equality. As Arrhenius has pointed out, "how valuable equal relations are might arguably depend on the level of welfare involved".[43] This is analogous to the commonly held belief among egalitarians that the negative value of unequal relations partly depends on the welfare levels of the involved individuals such that inequalities at low levels are worse than inequalities at high levels. Hence, it is open to positive egalitarians to give different values to equal relationships at different welfare levels such that, for example, no number of equal relationships at a very low positive welfare level is better in regard to egalitarian concerns than a certain number of equal relationships at a very high positive welfare level. This could, to take a simple example, be achieved by aggregating the value of equal relationships at high welfare levels linearly while the value of equal relationships at low positive welfare levels is aggregate by a strictly concave function with an upper limit.[44] Of course, there are more elaborate ways of developing this idea, for example, by having concave functions with different limits for each level of welfare. Hence, pace Segall, Positive Egalitarianism doesn't imply Repugnant Equality.[45]

Similar considerations hold for the following putative objection to Positive Egalitarianism discussed by Arrhenius:

---

[43] (Arrhenius 2013), p. 89.

[44] This is of course analogous to how one can avoid Parfit's Repugnant Conclusion even if one is committed to the positive contributive value of lives enjoying positive welfare. See e.g., (Arrhenius 2000a, 2005; Arrhenius, Ryberg, and Tännsjö 2014). Of course, any theory of this sort (and many other, e.g., Prioritarianism) faces the epistemic problem of how to determine the right shape of the concave function. This problem has no relevance for our discussion here, however.

[45] A broader version of Positive Egalitarianism according to which the egalitarian value is just an increasing function (instead of a strictly increasing function) of equal relations opens up another way of avoiding Repugnant Equality since it's compatible with a reachable upper limit to the total value of equal relations that has already been reached in A. Hence, this view would rank A and Z as equally good regarding egalitarians concerns. However, we don't find this view compelling.

> if one does assign some small positive value to such relations of equality, then one would have to accept an implication akin to the levelling down objection: Adding people with negative welfare to a population might make it in one respect better, although not all things considered better, since it might increase the value of equality in the population.[46]

As with Repugnant Equality and the levelling down objection, there are two possible answers here. One can bite the—in our minds just imaginary—bullet and accept a defeasible egalitarian reason to add equal people with negative welfare. Or, in extending Positive Egalitarianism to negative welfare levels, one can assign neutral value to equal relationships of negative welfare. In such case, it wouldn't follow that adding people with negative welfare to a population would make such population in one respect better. We prefer the first option since we are not persuaded by the levelling down objection and since we find it hard to motivate the second option.[47]

VII. Conclusion

Positive Egalitarianism is the view that ascribes positive value to relations of equality in addition to negative value to relations of inequality. In this paper, we have developed this view, argued that it is a distinctive view from the 'received' view on the value of equality, Negative Egalitarianism, and shown how some of the recently presented criticisms against Positive Egalitarianism are unconvincing.

In section II, we discussed a case of mere addition. Against Segall, we argued that the population that results from the mere addition of better off individuals can be judged as better than the original one by Positive Egalitarianism given that it contains many more relations of equality. This judgement, we also argued, will depend on what weight to give to bigger increases in relations of equality as opposed to smaller increases in relations of inequality.

In section III, we introduced and discussed the Progression Case. Segall raised this case aiming to show the alleged counterintuitive implications of Positive Egalitarianism. Contrary to Segall, we argued that it is not only Positive Egalitarianism that ranks population Z as better than Z' in the Progression Case. Negative Egalitarianism coincides with Positive Egalitarianism in this judgement. We provided a demonstration that supports the ranking of Z as better than Z' for both of these views.

---

[46](Arrhenius 2013, p. 89). It's repeated as an objection in (Segall 2016), p. 78.

[47] The motivating idea could be that equality only has value when it's equality of something good. This seems doubtful, however, since we seem to value equality among bad things too, such as carrying burdens equally or equal punishment for similar crimes.

In section IV, we discussed another argument that has been provided to support that Z' is better than Z with respect to egalitarian concerns, namely that Z contains a clear class division while Z' doesn't. Against Segall, we argued that given the broad notion of welfare we use, the welfare differences cannot be taken as proxies for class differences. And if they could, just because the grouping leading to class division is more obvious in Z than in Z', this doesn't imply that there wouldn't be ways in which the individuals in Z' could be also grouped to lead to a class division.

In section V, we responded to an argument posed by Temkin against Positive Egalitarianism that says that ascribing positive value to relations of equality is only significant if people can have the exact same amount of welfare, and given that this is often not the case, Positive Egalitarianism, although theoretically interesting, is of no practical relevance. To respond to this criticism, we provided a view called Rough Positive Egalitarianism. This view ascribes positive egalitarian value to relations of individuals that are roughly equal to each other which, we argue, fits the way people often think about the value of equality. We also defended Rough Positive Egalitarianism from the criticism that it entails the judgement that in the Progression Case, Z' is better than Z with respect to equality. Although Z' contains a lot of relations of rough equality since the differences in welfare between every adjacent individual are very small, this doesn't outweigh the fact that Z' also contains a great number of relations of inequality, and of a great size.

Finally, in section VI we argued that the so-called Repugnant Equality isn't clearly counterintuitive for egalitarians since this conclusion involves a claim just about betterness in regard to egalitarian concerns but not about all things considered betterness. Moreover, Repugnant Equality doesn't follow from Positive Egalitarianism since positive egalitarians can give different values to relations of equality depending on at which level these hold.

All in all, we think, the arguments so far presented against Positive Egalitarianism are unpersuasive and thus don't give us any decisive or new reasons to reject views that ascribe positive value to equality. Hence, it might still play an important role in population ethics and in the evaluation of future population outcomes, and as such, deserves further attention. Of course, there are many interesting unresolved issues that remain for positive egalitarians to address, to which we hope to return in future work.[48]

[48] We would like to thank Krister Bykvist, Tim Campbell, Nils Holtug, Kasper Lippert-Rasmussen, Shlomi Segall, Orri Stefánsson, and Folke Tersman for very helpful discussions. Thanks also to the audiences at: the Society for Applied Philosophy Conference, Copenhagen, June 2017; the workshop *Ethics and Political Philosophy*, University of Copenhagen, October 2017; the Institute for Futures Studies' PPE-seminar, September 2017; and the higher seminar at the Department of Philosophy, Uppsala University, October 2017, for useful questions and comments. Financial support from the Swedish Research Council and Riksbankens Jubileumsfond is gratefully acknowledged.

References

Arneson, Richard J. 1989. 'Equality and Equal Opportunity for Welfare'. *Philosophical Studies* 56 (1): 77–93.

Arrhenius, Gustaf. forthcoming. 'Population Ethics: The Challenge of Future Generations'. Oxford University Press.

———. 2000a. 'An Impossibility Theorem for Welfarist Axiologies'. *Economics and Philosophy* 16 (02): 247–266.

———. 2000b. *Future Generations: A Challenge for Moral Theory*. Uppsala: University Printers. http://www.diva-portal.org/smash/record.jsf?pid=diva2:170236.

———. 2005. 'Superiority in Value'. *Philosophical Studies* 123 (1/2): 97–114.

———. 2009. 'Egalitarianism and Population Change'. In *Intergenerational Justice,* edited by Axel Gosseries and Lukas Meyer, 1st Edition, 325–49. Oxford: Oxford University Press.

———. 2013. 'Egalitarian Concerns and Population Change'. In *Measurement and Ethical Evaluation of Health Inequalities,* edited by Ole Frithjof Norheim, 74–91. Oxford: Oxford University Press.

———. 2016. 'Inequality and Population Change'. Mimeo, Institute for Futures Studies. Stockholm.

Arrhenius, Gustaf, Jesper Ryberg, and Torbjörn Tännsjö. 2014. 'The Repugnant Conclusion'. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2014. http://plato.stanford.edu/archives/spr2014/entries/repugnant-conclusion/.

Barry, Brian. 1989. 'Utilitarianism and Preference Change'. *Utilitas* 1 (02): 278–282.

Bommier, A., B. Lanz, and S. Zuber. 2015. 'Models-as-Usual for Unusual Risks? On the Value of Catastrophic Climate Change'. *Journal of Envrionmental Economics and Management* 74: 1–22.

Bosmans, Kristof, Koen Decancq, and Andre Decoster. 2014. 'The Relativity of Decreasing Inequality Between Countries'. *Economica, LSE* 81 (322): 276–92.

Braybrooke, David. 1987. *Meeting Needs*. Studies in Moral, Political, and Legal Philosophy. Princeton, N.J.: Princeton Univ. Press.

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

———. 2004. *Weighing Lives*. Oxford: Oxford University Press.

Bykvist, Krister. 1998. *Changing Preferences: A Study in Preferentialism*. Uppsala: Acta Universitatis Upsaliensis.

Cohen, Gerald A. 1989. 'On the Currency of Egalitarian Justice'. *Ethics* 99 (4): 906–944.

———. 1993. 'Equality or What? On Welfare, Goods and Capabilities'. In *The Quality of Life*, edited by Amartya Sen and Martha Craven Nussbaum. WIDER Studies in Development Economics, 99-0851294-3. Oxford: Clarendon Press.

Dworkin, Ronald. 1981a. 'What Is Equality? Part 1: Equality of Welfare'. *Philosophy & Public Affairs* 10 (3): 185–246.

———. 1981b. 'What Is Equality? Part 2: Equality of Resources'. *Philosophy & Public Affairs* 10 (4): 283–345.

———. 2000. *Sovereign Virtue: The Theory and Practice of Equality*. Cambridge, Mass: Harvard University Press.

Feldman, Fred. 1997. *Utilitarianism, Hedonism, and Desert: Essays in Moral Philosophy*. Cambridge Studies in Philosophy. Cambridge: Cambridge University Press.

———. 2004. Pleasure and the Good Life: Concerning the Nature, Varieties, and Plausibility of Hedonism. Oxford: Oxford University Press.

Geruso, Michael, and Dean Spears. 2018. 'Heat, Humidity, and Infant Mortality in the Developing World'. *IZA - Institute of Labor Economics* IZA DP No. 11717.

Griffin, James. 1990. *Well-Being: Its Meaning, Measurement, and Moral Importance*. Oxford: Clarendon Press.

Hales, S., S. Kovats, S. Lloyd, and D. Campbell-Lendrum. 2014. *Quantitative Risk Assessment of the Effects of Climate Change on Selected Causes of Death, 2030s and 2050s*. Geneva: WHO.

Hare, Richard M. 1981. *Moral Thinking: Its Levels, Method, and Point*. Oxford: Clarendon.

Hurka, Thomas. 1993. *Perfectionism*. Oxford Ethics Series. New York: Oxford University Press.

Kawchuk, Risa. 1996. *(In)Equality*. MPhil-Thesis. Dept. of Philosophy, University of Calgary.

Mejean, Aurelie, Antonin Pottier, Setephane Zuber, and Marc Fleurbaey. 2017. 'Intergenerational Equity under Catastrophic Climate Change.' *Documents de Travail Du Centre d'Economie de La Sorbonne*.

Mosquera, Julia. 2017a. 'An Egalitarian Argument against Reducing Deprivation'. *Ethical Theory Moral Practice* 20 (5): 957–968.

———. 2017b. 'Disability, Equality, and Future Generations'. PhD-thesis. University of Reading.

Nielsen, Kai. 1996. 'Radical Egalitarianism Revisited: On Going Beyond The Difference Principle'. *Windsor Yearbook of Access to Justice* 15: 121–58.

Parfit, Derek. 1984. *Reasons and Persons*. 1991st ed. Oxford: Clarendon.

———. 1986. 'Overpopulation and the Quality of Life'. In *Applied Ethics*, edited by Peter Singer, 1 edition, 145–64. Oxford: New York: Oxford University Press.

Persson, Ingmar. 2001. 'Equality, Priority and Person-Affecting Value'. *Ethical Theory and Moral Practice* 4 (1): 23–39.

Rabinowicz, Wlodek. 2003. 'The Size of Inequality and Its Badness: Some Reflections Around Temkin's Inequality'. *Theoria* 69 (1–2): 60–84.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, Mass.: Belknap Press.

Segall, Shlomi. 2016. *Why Inequality Matters: Luck Egalitarianism, Its Meaning and Value*. Cambridge: Cambridge University Press.

Sen, Amartya. 1980. 'Equality of What?' In *The Tanner Lectures on Human Values,* edited by S. McMurrin. Cambridge: Cambridge University Press.

———. 1992a. *Inequality Re-Examined*. Reprint edition. New York: Harvard University Press.

———. 1992b. 'Justice and Capability'. In *Inequality Reexamined*, 73–101. New York: Russell Sage Foundation; Harvard University Press.

———. 1993. 'Capability and Well-Being'. In *The Quality of Life*, edited by Amartya Sen and Martha Craven Nussbaum. WIDER Studies in Development Economics. Oxford: Clarendon Press.

Sumner, Leonard Wayne. 1996. *Welfare, Happiness, and Ethics*. New York: Clarendon Press.

T. Carleton et al. 2018. 'Valuing the Global Mortality Consequences of Climate Change Accounting for Adaptation Costs and Benefits'. *Becker Friedman Institute* Working Paper No 2018-51.

Tännsjö, Torbjörn. 1998. 'Utilitarianism and Common Sense Morality'. In *Hedonistic Utilitarianism*, 153–71. Edinburgh: Edinburgh University Press.

Temkin, Larry S. 1993. *Inequality*. Oxford Ethics Series. New York: Oxford University Press.

Stéphane Zuber[1] & Marc Fleurbaey[2]

Discounting and Intergenerational Ethics

The question of social discounting is central in intertemporal cost-benefit analysis that often shapes economists' recommendations regarding climate policy. The practice of discounting has been the object of heated debates among economists and philosophers, revolving around the issue of intergenerational ethics. In this chapter, we review the different arguments for and against specific values of social discounting. We show that there are actually two different ethical issues at stake: 1) the question of impartiality (or equal treatment of all generations); 2) the question of priority to the worse-off (aversion to inequality in resources, capabilities or welfare). These questions have emerged in the utilitarian approach and can be neatly separated in that case. They also have very different consequences for climate policy. We then argue that the question of social discounting is not confined to the utilitarian framework as it more generally describes the social value of income (or capability or welfare) transfers to future generations. Lastly, we discuss the many limitations of social discounting as a tool for policy analysis.

[1] Paris School of Economics – CNRS, France. Address: Centre d'Economie de la Sorbonne, 106-112 boulevard de l'Hôpital, 75013 Paris, France. E-mail: Stephane.Zuber@univ-paris1.fr.
[2] Woodrow Wilson School and University Center for Human Values, Princeton University, USA. E-mail: mfleurba@princeton.edu.

Introduction

The social discount rate is an important tool in intertemporal cost-benefit analyses used by economists to inform long-term policy choices, for instance mitigation policies aiming at reducing greenhouse gas emissions. The social discount rate is used to convert future monetary costs and benefits in present value. For instance, a rate of 2 % (=0,02) means that a project whose cost is $1,000,000 in 50 years has a present value of $1,000,000 / $(1+0,02)^{50} \approx$ $371,528 today. As illustrated by this simple example, even low discount rates can significantly reduce the value of future costs and benefits. Moreover, if the social discount rate is 3% the present value of the same project would be about $228,107, that is, two thirds of a the present value using the 2% rate. Hence, apparently small changes in the social discount rate have huge implications in terms of how we value future economic impacts.

As acknowledged in the last report of Working Group 3 of the IPCC (Intergovernmental Panel on Climate Change), "the use of a temporal discount rate has a crucial impact on the evaluation of mitigation policies and measures" (Kolstad et al. 2014, p. 211). The Working Group asserts that the appropriate risk-free social discount rate should be between one and three times the anticipated growth rate in real per capita GDP, but also underscores that these values are based on the so-called Ramsey rule and that ultimately there are normative choices to be made. In this chapter, we aim at presenting some of these normative choices. We also argue that one does not need to restrict attention to the Ramsey rule and that the discount rate is a general tool that can be applied in many ethical frameworks.

Heated discussions around specific values of the social discount rate have aroused after the publication of the Stern review's recommending strong action to reduce greenhouse gas emissions (Stern 2006). A key principle that was discussed was intergenerational equity and the discussion mostly relied on the Ramsey rule. In Section 1 of this chapter, we present the Ramsey rule and explain how it encompasses two different dimensions of equity: impartiality (the equal treatment of all generations) and preference for more equal distributions of resources. We explain why these two aspects can go in opposite directions. We also present the different positions in the debate that followed the Stern review.

The Ramsey rule emerged in the Utilitarian moral framework. In his seminal paper, Ramsey (1928) explicitly assumed that "enjoyments and sacrifices at different times can be calculated independently and added" (Ramsey, 1928, p. 543) and used a criterion that explicitly adds utilities across periods. An important question then could be whether social discounting is tied to an excessively narrow (e.g., utilitarian, economistic) ethical approach, or whether it can accommodate a variety of relevant principles and values. One thesis defended in this chapter is that,

while usual practice of discounting is indeed restricted to an unduly narrow ethical framework, the methodology itself is rather flexible. Still, it is part of a cost-benefit approach, which relies on a consequentialist axiology. Section 2 provides a general ethical framework in which discount rates may be obtained and we argue that this framework can encompass many moral views about who matters, how claims should be balanced, how advantages should be distributed, and how to measure individual advantage.

Despite its salience, the issue of discounting cannot be considered as the only important one in the evaluation of policies that affect future people (for instance, climate policies). We briefly recall why discount rates are theoretically relevant only for the analysis of marginal changes and can therefore not be used to assess abrupt regime shifts or existential risks. They are also not the right tool to use for choice that may affect the size or composition of the future population. Last, the discounting issue should also not obscure other important equity issues like the distribution of cost and benefits within generations. We present these limitations of social discounting in Section 3.

1 Discounting: definition in the utilitarian framework and equity issues

As explained in the introduction, the standard approach to determining the social discount rate is based on the *Ramsey equation*. The perspective adopted to derive this equation is that of a benevolent social planner (or sometimes a representative agent) that seeks to maximize a value function. Such a value function assigns a real number to each possible state of affairs, with better states of affairs being assigned a greater number. The standard approach typically describes states of affairs as streams of consumption for all future generations, with c_t the consumption of generation t (where $t = 0$ denotes the current generation), and the value function is given by[3]

$$V(c) = \sum_{t=0}^{\infty} e^{-\delta t}\, u(c_t)\,, \qquad (1)$$

where function u is the utility function (transforming consumption levels into utility numbers) and δ is the so-called `*utility discount rate*' or `rate of pure time

[3] We do not account for population size in this formula to simplify the exposition. We could more generally write utility as $u(N_t, c_t)$, with N_t the population size in period t. Typically, the total utilitarian approach takes $u(N_t, c_t) = N_t v\left({}^{c_t}/_{N_t}\right)$.

preference'. The term $e^{-\delta t}$ can be seen as a decreasing weight put on the utility of future generations.

Equation (1) is often further simplified by assuming that $u(c) = c^{1-\eta}/(1-\eta)$. In that case, the η parameter measures how rapidly marginal utility decreases when consumption increases (the formula thus accepts that marginal utility is decreasing, a standard assumption in the utilitarian tradition).

The idea of social discounting consists in measuring the value of a small increase in consumption in period *t* as equal to the value of a small increase in consumption today, discounted by a factor $1/(1+\rho_t)^t$. The ρ_t parameter is the social discount rate. From the value function described in Equation (1), it can be derived from the Ramsey equation:

$$\rho_t = \delta + \eta g_t, \tag{2}$$

where g_t is the average growth rate of consumption between the current period and period *t*.

The Ramsey Equation (2) clearly distinguishes two reasons for discounting. One is pure time discounting expressed by parameter δ: we discount future consumption or future damages because we discount the welfare of future people. This first reason has given rise to a debate about intergenerational equity that could better be expressed as a debate about impartiality. The second part of the Ramsey equation combines the elasticity parameter η and the growth rate of consumption. This is "discounting for growth": given that future generations are richer when g_t is positive, their consumption has less priority. The rate of decrease in the value of future consumption is proportional to growth, and the proportionality parameter η represents the strength of the redistributive motive. A higher η represents a higher willingness to reduce consumption inequalities.

We now separately discuss these two rationales for discounting.

1.1 Discounting and intergenerational impartiality: "normative" and "positive" approaches

The main controversy in the economics of climate change was probably the Nordhaus-Stern debate that arose after the publication of the Stern review (Stern, 2006). The Stern review promoted a strong action to reduce greenhouse gas emissions, while Nordhaus promoted a gradual response (Nordhaus, 2008). As explained by Nordhaus himself (Nordhaus, 2007), a big part of the difference lies in using different values of the social discount rate: the Stern review is based on a 1.4% discount rate, while Nordhaus preferred a 5.5% discount rate. Part of the difference

is due to using a different value of the utility discount rate: Stern argued in favor of $\delta = 0.1\%$ while Nordhaus chose $\delta = 1.5\%$.

The line of argument used by Stern to justify his low value of the utility discount rate is based on the principle of impartiality. This was hardly a new line of argument: it can be traced back to Sidgwick who argued that "[...] the time at which a man exists cannot affect the value of his happiness from a universal point of view" (Sidgwick, 1907, p. 414). Ramsey himself used the idea of impartiality to justify a zero-utility discount rate. In the end, Stern does not use a zero-utility discount rate but a very low rate to account for a small probability that future generations may not exist.[4] The impartiality argument is the main argument in favor of a zero or near-zero utility discount rate.[5]

Despite this strong impartiality argument, several authors have insisted that several reasons may explain a positive pure time discount rate. A first line of argument in favor of a positive utility discount rate was provided by Koopmans (1960) who produced an influential axiomatization of discounted utilitarianism based on the Pareto principle combined with rationality and parsimony principles of time consistency and invariance of social evaluation. From this initial contribution stemmed a very rich (but technical) literature showing the incompatibility between the Pareto and impartiality principles when one considers an infinite sequence of successive generations (Diamond, 1965; Basu and Mitra, 2003; Zame, 2007; Lauwers, 2010).

One possible conclusion from this strand of literature would be that we cannot maintain the impartiality requirement in the intergenerational context because the Pareto principle is deemed to have more normative appeal. There are however several objections. A first objection is that it does not seem plausible that there will be infinitely many future generations, at least if we only consider generations of humans on Earth (there may exist other living species in other galaxies or worlds, but one may dispute whether they are morally relevant for us). A key problem then is that we do not know how many future generations there will be: we cannot plausibly specify a number of years until humans disappear. A well-known remedy has been proposed: to introduce an extinction (or existential) risk (Dasgupta and Heal, 1979; Stern, 2006). In the utilitarian case, this risk provides a foundation for a utility discount rate δ equal to the hazard rate of extinction as explained in footnote

[4] With this interpretation, the utility discount factor $e^{-\delta t}$ is the probability that the future generation of period t exists.

[5] Greaves (2017, p. 405) mentions another argument based on an extended Pareto principle. We believe that this argument, which involves a choice made an individual between existing today with some consumption level or living tomorrow with this same level, is not real realistic unless we consider a veil of ignorance context that actually serves to promote the impartiality principle.

2 (for a more general treatment, beyond the utilitarian case, see Fleurbaey and Zuber, 2015b).

Another objection is that violations of impartiality occur only because we want to have a value function that is able to rank and compare all situations. We may consider a less demanding moral theory based on *incomplete* criteria (that may not always say which situation is morally better). Several incomplete versions of utilitarianism – without utility discounting – have been proposed, in particular the overtaking criterion (Von Weizsäcker, 1965) and the Gale criterion (Gale, 1967). Last, let us underline that Koopmans' (1960) argument in favor of a positive value of δ does not provide any guidance about the exact value of this parameter. In principle, it could be as low as one wishes so that, for any practical purpose, we may just neglect this part of the Ramsey equation.

In another vein, Arrow (1999) argued that the present bias introduced by utility discounting is not only a mathematical necessity, related to the infinite horizon framework, but is also ethically justified, on the grounds that it reflects a permissible agent-relative preference for ourselves and our own projects. A similar form of agent relative morality was defended by Dasgupta (2016) who proposed a form of generation-relative utilitarianism. One key intuition developed by Arrow and Dasgupta in favor of a large enough utility discount rate is that it is not morally acceptable to demand excessively high savings rates of any one generation: simple growth models with a value of δ close to zero typically imply a very large savings rate (see Mirrlees, 1967). This drawback of undiscounted utilitarianism was already mentioned by John Rawls who declared that "the utilitarian doctrine may direct us to demand heavy sacrifices of the poorer generations for the sake of greater advantages for the later ones that are far better off" (Rawls, 1971, p. 253). He went on to say that "these consequences can be to some degree corrected by discounting the welfare of those living in the future" (Rawls, 1971, p. 262). However, Rawls never argued in favor of discounted utilitarianism but simply wanted to point out a flaw of utilitarianism in the intergenerational context. The concern that utilitarianism may demand too large sacrifices from the current generation may also not be real: what matters for optimal savings is the whole consumption discount rate ρ_t, not simply the utility discount rate: reasonable levels of investments can be obtained in the undiscounted utilitarian framework if one chooses large enough values of η (see Asheim and Buchholz, 2003).

While the two previous lines of arguments against the zero-utility discount rate may be related to normative considerations, many economists have preferred to offer reasons that are not directly stated as ethical reasons. Some scholars have labelled approaches relying on these reasons as "descriptive" approaches (see Arrow et al., 1996) or "positivist" approaches (Posner and Weisbach, 2000). These

approaches mainly use a revealed preference argument.[6] Most people do in fact discount their future utility, as revealed for instance in market interest rates. Given that collective actions should be selected on the basis of aggregating individual preferences, a utility discount rate should reflect people's present bias. In particular, Nordhaus (2007) famously declared Stern's approach as undemocratic and depicted it as a situation where a utilitarian elite (epitomizing "the dying embers of the British empire") makes decisions based on its own rather than the population's belief. Several objections can be made to the revealed preference argument. First, even if markets do aggregate preferences in some way (provided markets are well-functioning) they do so in a very specific way that may not be democratic. Indeed, the aggregation depends only the preferences of those people who are active on the market and on their initial wealth, so that poorer people preferences are typically not represented. Furthermore, future people's interests and preferences are not represented (at least not directly: they may be partially represented only to the extent that current people care about them). Hence, even if the descriptive approaches do not explicitly take an ethical stance on how advantages should be distributed across generations, they do implicitly rely on ethical assumptions. These assumptions are broadly that only current generations, and among them mostly the wealthier people or at least those who are active on markets, may have a say on how to allocate goods between periods, even in the long term.

The many objections to arguments in favor of a strictly positive utility discount rate explain why the authors of the 5th Assessment report of the IPCC mention a "relative consensus in favor of $\delta = 0$" (Kolstad et al. 2014, p. 230).

1.2 Discounting and aversion to intergenerational inequalities

The second part of the Ramsey equation (2) has to do with the fact that future generations may be richer. It is the product of the growth rate of consumption, which is clearly an empirical quantity (albeit a very uncertain one), with the elasticity parameter η. In the economic literature and in most presentations of the Ramsey rule, three main interpretations of this parameter have been offered (see for instance Greaves, 2017). It may represent:

[6] Posner and Weisbach (2000) have a different "positivist" line of argument to choose the social discount rate (without reference to the Ramsey equation): an opportunity cost argument. We should never choose projects with returns lower than the market interest rate because otherwise we would incur a loss – we could have invested in a market fund and get a larger benefit in the future. But the identity between the social discount rate and the market rate is justified only in the very specific case where we are already at the economic optimum (so that no more investments are required). It also assumes that the maturity of the market investment is the same as that of the proposed policy, which is unlikely for very long term policies like climate policy. We have discussed this line of arguments in more details in Fleurbaey and Zuber (2013).

- Individuals' relative risk aversion;
- Individuals' inverse elasticity of intertemporal substitution;
- Aversion to inequality.

The choice of one of these interpretations is consequential. As highlighted in Atkinson et al. (2009), empirical estimates of these three quantities are usually very different, which may explain the very wide range of value found in the literature (from 1 to 3 or 4 according to Kolstad et al. 2014, p. 230). Although the economic literature mentions these three interpretations, it mainly presents them as three empirical strategies to calibrate η rather than appealing to normative reasons to choose one of them. We would like to argue that they correspond to specific interpretations of the utilitarian formula (1), including non-utilitarian ones.

The two first interpretations (in terms of individuals' risk or temporal preferences) correspond to a specific view on utility, namely utility as preference satisfaction. Although this is the standard interpretation of utility in economic theory, this is not uncontroversial for utilitarians that may prefer a hedonistic interpretation, or definitions appealing to people' judgement on their own life. But even if one accepts the preference satisfaction definition of utility, it appears that one has to choose what preferences are relevant: risk preferences or instantaneous temporal preferences.

The use of risk preferences to measure utility has a long history in economics. Harsanyi (1953, 1953) famously provided two frameworks to justify a representation of utility by risk preferences. Harsanyi's (1953) *impartial observer theorem* assumes that, behind a veil of ignorance, individuals have extended preferences on prospects of outcomes and identities, so that they can compare welfare across outcomes and preferences. Harsanyi assumed that all deliberators behind the veil of ignorance will have the same extended preferences that are revealed by choices under uncertainty. Harsanyi's (1955) *aggregation theorem*, which involves the Pareto principle applied to risky situations, characterizes utilitarianism as a sum of Von Neumann-Morgenstern individual utilities.[7] Although they have attracted less attention, time preferences can give a foundation for individual utility by arguments similar to those in Harsanyi's (1955). Zuber (2011, Prop. 1) showed that when individual preference are time separable, social aggregation satisfying the Pareto principle must a sum of these utilities.

[7] Sen (1976) and Weymark (1991) argued that, while Harsanyi's theorem establishes that social welfare is a sum of VNM utilities, it does not follow that it is the sum of individuals' welfare levels. But this is beside the point of the theorem, whose potent message is that social evaluation must rely on a (weighted) sum of VNM utilities. See Fleurbaey and Mongin (2016) for further discussion.

Interestingly, both time and risk preferences have been used to construct welfare-like measures for the case of health. Indeed, QALYs (Quality Adjusted Life Years) that are widely used health measures are often calibrated using two methods (Drummond, Stoddart and Torrance, 1987): the standard gamble methodology (using risk preferences) and the time trade-off methodology (using time preferences).

The other prominent interpretation of the elasticity parameter η is that it controls social attitudes towards inequality (it is the coefficient of relative inequality aversion in the terminology used in economics). The foundation here may not be in terms of decreasing marginal utility of consumption, as in the standard utilitarian approach, but in terms of social priority for consumption of poorer people. Most empirical approaches to estimate this inequality aversion parameter are based on individuals' attitudes towards redistribution in stated-preference experiments (see, e.g., Atkinson et al. 2009) or on actual redistribution policies (see, e.g., Tol 2010). A method providing intuition about the inequality aversion parameter is the leaky bucket thought experiment (Okun 1975), where one has to declare how much one is willing to lose in a transfer of money from richer to poorer people. But some authors argue that in the intergenerational context, we should rather ground our moral intuitions in experiments concerning the saving rate (Dasgupta 2008).

In all cases, the parameter will control the marginal social value of consumption for any individual, depending on her initial consumption level. A larger value of η implies a greater priority for poorer people or poorer generations. Thus, a value function exhibiting "greater concerns for intergenerational inequality" (in the specific sense of inequality aversion with respect to consumption) typically exhibits a larger social discount rate.

We then end up with two similarly opposite effects of ethical principles of justice on the social discount rate. The principle of impartiality, discussed in 1.1, would require a very low level of δ (actually $\delta = 0$ if we do not account for a risk of extinction) and thus a low level of the discount rate. A principle of equality (or aversion to inequality), as applied to consumption levels in the discussion above, would require a high level of η and thus a high level of the discount rate, at least in the standard case where the consumption growth rate g_t is positive.[8] Of course, there is no direct logical connection between the principles of impartiality and

[8] See Equation (2). Of course, if $g_t < 0$, i.e., when future generations are poorer than current generations, more aversion to consumption inequality will decrease the social discount rate, possibly yielding a negative discount rate. This will also happen in a framework with uncertainty about future growth, provided that it is sufficiently likely that situations where $g_t < 0$ may happen (see Fleurbaey and Zuber, 2013).

equality, but one may expect that foundations of the social discount rate that are based on the idea of intergenerational justice will promote both a low value of δ and a high value of η.[9] All in all, one can conclude that there is no clear reason why intergenerational justice in general would promote high or low values of the social discount rate.

2 Discounting beyond utilitarianism

Discounting is not restricted to Utilitarianism. However, it still entails making strong ethical assumptions. Indeed, any notion of discounting is derived from the existence of a value function, whose general form can be as follows:

$$V = F(c_0, c_1, \dots, c_t, \dots), \tag{3}$$

where c_t is a vector of all "goods" that matter in period t and that can be discounted to compare with the value of good(s) in period 0.

The existence of a value function means that we are trying to assess and compare situations in terms of their goodness or, more accurately, their betterness. This is in contrast with ethical frameworks pertaining to the theory of the right or other approaches involving notions of harm, virtues or duties. Also, it means that we are primarily concerned with outcomes or consequences of actions. Therefore approaches for which discounting can be a relevant tool belongs to a broad class of maximizing consequentialist theories. Note however that some people who don't think that consequentialism is the only relevant ethical view may still think that consequences are part of the ethical considerations we should rely on. For instance, Broome (2012) argues public morality can focus on the pursuit of goodness while individuals can focus on avoiding actions that harm other (future) people, including by compensating potential harms through carbon offsetting. This division of labor can be debated but cost benefit analysis and thus discounting are important tool for coordinating the pursuit of the common good and thus pertain to axiological public morality. More generally, discounting may be relevant in an ethical theory that is not purely consequentialist: it is relevant only in so far as consequences are relevant.

Equation (3) is a very broad and general definition of the value function. It may include several "goods" or ethical dimensions that may range from human-centered individualistic and materialistic considerations (the amount and distribution of personal consumption goods in a population in a given period) to more holistic and

[9] And indeed, for instance, Dasgupta (2008) criticized the Stern report for using principles of intergenerational justice to justify a low value of δ without considering such principles when setting the value of η.

non-speciesist (we may include dimensions like the quality of social relations, the level of biodiversity and protection of other species in the list of goods). The key assumption made by Equation (3) is that these different goods are measurable as well as comparable: we can derive overall good from them (and thus implicitly trade-off the different dimensions).

With the general value function described in Equation (3), we cannot obtain a precise description of the main elements of the social discount rate. We will thus focus on a more specific value function encompassing many individualistic consequentialist ethical theories.

2.1 A general formula for individualistic consequentialist ethics

The more specific class of value functions we will focus on is described by the following formula:

$$V = W\left(\left(w(i,c_i)\right)_{i\in N_0}, \left(w(i,c_i)\right)_{i\in N_1}, \dots, \left(w(i,c_i)\right)_{i\in N_t}, \dots\right), \qquad (4)$$

where each c_i is a vector of "goods" available to person *i* and each person *i* belongs to a specific generation (N_t is the set of all individuals leaving in generation *t*; typically generation 0 is the current generation, but we may imagine that some past generation could be the "first" generation).[10] Function *w* is an individual advantage function that depends on the identity of each individual (therefore *w* depends on *i*) and on the consumption vector.

Formula (4) provides a very flexible framework to encompass many ethical theories. Different theories can be described as taking a stance on three different issues: the scope of justice (the population N_t of individuals included in each generation); the currency of justice (the individual advantage functions *w* and the goods included in the vectors c_i of personal goods); the shape of justice (the "aggregator function" *W* that combines and weighs the advantages of the different individuals).

The scope of justice: The question is to decide which entities are the legitimate recipients of burdens and benefits. Formally, in Equation (4) the question is to decide who is included in population N_t in each period. We may even ask whether population N_t should appear for $t > 0$. As suggested in "descriptive" approaches to

[10] There is an ambiguity in the economic literature with the notion of generation: usually economic models consider only time periods but name such periods a "generation". Most ethical theories would use indices of individual lifetime well-being to weigh the claims of different people, while in most applications only indices of momentary well-being appear. See Greaves (2017, p. 396) for a discussion of this matter.

social discounting, some may argue that only current generations can have legitimate claims and that future generations claim are taken into account only insofar as some people in current generation care about them. However, most approaches to climate ethics would include both current and future generations.

Within a period, one may then wonder whether we should include all persons irrespective of the country they belong to. Schelling (1995) famously suggested distance in space might justify different treatment of individuals and that we may care less about people in the far away countries. Again, the impartiality principle seems to prevent us for making a difference between individuals on the basis of where and when they live. But the literature about global justice and cosmopolitanism has produced (controversial) arguments to justify some bias against aliens in defining social priorities (Rawls 1999; Nagel 2005).

Then we could also argue that the scope of justice includes not only humans but also nonhuman animals or even other nonhuman species. For instance, climate change is an important stressor for biodiversity so that it may overwhelm species that are slow to move or adapt. Of course, the instrumental value for humans of the environment may be included in the vector of goods c_i that is available to a person. But this completely ignores any possible intrinsic value. The literature on animal ethics and climate change is developing quickly (see Hsiung and Sunstein 2007, McShane forthcoming, Sebo forthcoming) and emphasizes the need to broaden the scope of justice. One of the key difficulties remains to identify principles of cross-species comparisons: how can we trade off human against non-human interests? Even within a hedonistic utilitarian approach, the question is not easy to settle: it is not at all obvious how to compare emotions across differently structured brains, or neural systems more generally.

In principle formula (4) could cover any scope discussed above. As highlighted by the problem of nonhuman species, it is however not simple in practice to extend the scope of justice as far as one would like and most applications restrict attention to the human population in all countries and all present and future periods.

The currency of justice: The question is to define what should be distributed and how the situation of the different people composing the population should be assessed and compared. In the philosophical and economic literature on social justice, many answers have been provided. Of course, one classical answer that we have already discussed before is that the function w in Equation (4) is an individual advantage function, as used in Utilitarianism. But even then, as emphasized before, there are several approaches to such an individual advantage function, including preference-satisfaction, extended-preferences, hedonistic or other mental-state approaches. One can give the generic label of welfarism to approaches relying on such individual advantage functions.

Several alternatives to welfarism have been proposed at least since the seminal book by Rawls (1971). Rawls proposed to replace welfare metrics with indices of primary goods: in that case, the vector c_i would be a vector of primary goods and $w(i, c_i)$ the corresponding index. Sen (1985) proposed the concept of capabilities reflecting the freedom or ability to achieve valuable functionings (i.e. "beings and doings"). In that case, the vector c_i should be not thought of as a vector of commodities but as access to functionings and the function w becomes an index of capabilities as developed in the economic literature (see Alkire 2016 for a recent review). Roemer (1996) proposed to develop indices of opportunities that should be equalized to achieve social justice. Opportunities are distributions of outcomes or advantages that people may choose or achieve through effort. Fleurbaey and Maniquet (2011) and Fleurbaey and Blanchet (2013) recently revived theories of equivalent-income to combine multiple dimensions of human achievement and welfare into a single index formally similar to function $w(i, c_i)$. There are many other possible views on the appropriate currency, including objective goods approaches (see Adler and Fleurbaey 2016 for presentations and comparisons of many different approaches).

Formally, any of these approaches could give rise to a specific application of the social discounting methodologies. There are however practical restrictions depending on the specific case. First, only goods that are included in the vector c_i can be discounted and converted into some corresponding present value. Some of the approaches we have discussed may include non-material goods (or even no material goods at all), which raises the question of how such non-material goods should be measured. Some of the approaches also consider opportunities described as menus or distributions of outcomes or achievements. What we should be discounting, then, are changes in those distributions, which is not mathematically as straightforward as discounting changes in simple quantities. But distributions can be seen as risky outcomes and a whole methodology has been developed to discount risky outcomes.[11]

The shape of justice: The question concerns the criteria to use to determine how to weigh the benefits accruing to different people. Beside the additive formula of Utilitarianism exhibited in Equation (2), where (weighted) welfare numbers are simple added, many other forms for the "aggregator function" *W* have been proposed and studied. A prominent alternative defended by Parfit (1997) and Broome (2004) is an additively separable formula that yields Prioritarianism. A general formula would be:

[11] We cannot extensively cover the case of risky outcomes in that chapter. The question has been discussed in depth in other papers, for instance Fleurbaey and Zuber (2015a), Greaves (2017) and Fleurbaey et al. (2019).

$$V = \sum_{t=0}^{\infty} \sum_{i \in N_t} \phi\big(w(i, c_i)\big), \quad (5)$$

With Prioritarianism, welfare numbers are transformed using a concave function ϕ that grants more priority to welfare that accrues to worse-off people compared to that accruing to better-off people. Such a formula has been applied to social discounting and climate policy (Fleurbaey and Zuber 2015b, Adler et al. 2017).

Another prominent option is Egalitarianism either in its strict form proposed by Temkin (1993) or in the modified form of maximin or leximin as suggested by Rawls (1971). They have usually been considered too extreme (including by Rawls himself, as recalled earlier) for application to intergenerational equity. Strict egalitarianism faces the levelling-down objection that we may prefer to reduce the welfare of everyone in society to promote equality and therefore has not been used by economists who seek efficient allocations. The maximin approach has counter-intuitive implications in the case of climate change policy: given that it only focuses on the worst-off, it completely discount impacts or outcome changes of all other individuals and thus, in all existing models, of all future generations. Given that the current worst-off generation pays for the cost of climate policy, we do not want to make any sacrifice for the sake of the future.

Sufficientarianism is the doctrine that the notion of sufficiency, understood as having a decent (or good enough) life, should be the key consideration for distributive justice. A version of sufficientarianism holds that as many people as possible should enjoy conditions of life that place them above a sufficiency threshold (Frankfurt 1987, 2000). Another version holds that we should give greater priority to helping worst-off persons up to the point at which these persons attain a good enough quality of life, but otherwise we should only maximize total welfare (Crisp 2003).

The economic literature has also provided several social criteria to aggregate individual welfare or advantage with the idea to promote a notion of sustainability. Chichilnisky (1996) proposed sustainable social preferences that combine a discounted sum of utilities and a long-run value. Asheim, Mitra and Tuggoden (2012) introduced a sustainable discounted utilitarian criterion similar to discounted utilitarianism in the sustainable case where future generations are better-off than the current generation, but which is similar to Maximin case for unsustainable paths. Zuber and Asheim (2012) have introduced a rank-dependent model that implies a relative priority (in contrast to the absolute priority of prioritarian criteria) to worst-off people. The model shares some similarities with discounted utilitarianism except that the pure-time discount rate is conceived as a social weight prioritizing the interests of least-advantaged people.

The W function is therefore very flexible, so that the discounting technique can be applied to many views regarding how interests should be balanced. Most consequentialist approaches would fit into the model proposed by Equation (4).

2.2 The main elements of the discounting formula

In Fleurbaey and Zuber (2015a), we developed a general methodology to compute a social discount rate for the general value function described in Equation (4). The social discount rate then generally represents the rate of change in the value of a specific (or composite) good in the future period t compared to a reference (composite) good in the current period.[12] This rate will depend on two key elements: the rate of change in marginal advantage derived from the consumption of the good; the rate of change in the social priority of individual advantage. The "marginal advantage derived from the consumption of the good" is similar to the concept of marginal utility of consumption. The only difference is that the individual advantage function $w(i, c_i)$ is not necessarily a utility function that measures changes in pleasure or happiness, but may measure changes in opportunities, capacities or other concepts.

What we call the "social priority of individual advantage" is a generalized notion of priority that does not necessarily correspond to the concept developed in the prioritarian theory.[13] It simply measures the social or ethical marginal value of advantage for a specific person. This makes it possible to compare and balance the distributive claims or needs of different people: is such or such increase in the advantage of the current generation more or less valuable than such or such increase in the advantage of a generation living in one thousand years?

In a utilitarian formula, social priority of utility would be the same for all individuals in all generations. In a discounted utilitarian formula, this social priority would be $e^{-\delta t}$ for a person of generation t, and thus decreasing through time. In a prioritarian formula, this social priority would be lower the better-off a person is. In a sufficientarian formula, the social priority would be greater for badly-off individuals up to some level of advantage and then the same for all. In an egalitarian maximin formula, all individuals would have zero priority except the worst-off persons, implying an extreme discounting formula.[14]

---

[12] The term "specific good" means that we consider a particular good in the vector c_i for a specific individual $i \in N_t$. The term "composite good" means that we consider an aggregate or equivalent quantity to represent the level of all goods either for an individual or the society at large in a given period. Typically, economic models focus on average consumption in a population.

[13] See Parfit (1997) for the introduction of this notion of priority and Broome (2015) for a defense of the concept.

[14] For instance, the maximin case with one person per generation and one good, the discount rate would be infinitely large if the current generation is the worst-off generation.

The simple decomposition in terms of marginal advantage and social priority of individual advantage can actually become more complex in practice for three reasons: 1) there are multiple goods; 2) there are several people in a generation; 3) there may be some risk or uncertainty about future outcomes.

Contrary to the simple case of formula (1) and the associated Ramsey equation (2), the value function (4) allows the advantage measure for individuals to depend on several goods (or resources or attributes). These different goods will generally have different discount rates, reflecting future changes in their relative value or price. One way out of this complication is to construct an "aggregate" good that can be discounted with a single discount rate. But, as explained in Greaves (2017), we must then be careful not to forget the issue of changing relative prices: if some good (for instance environmental quality, the level of biodiversity, etc.) becomes less abundant in the future compared to another good (for instance material consumption), its relative price will increase making it more valuable. And more to the point: not only the relative market price may change, but the relative value for the ethical assessment may change (where these relative value is computed as a *shadow price* using the value function in Equation (4), which may not be reflected in market prices). Retaining good-specific discount rates (see for instance Gollier 2010), on the other hand, is more transparent and makes it possible to highlight the importance of certain goods when assessing policies affecting several generations: for instance, the "ecological discount rate" for future damages on the environment may become negative if we think that environmental quality will decrease in the future (Gollier 2010).

Future generations are composed of several individuals that may be affected by the decisions we take today. For ethical assessment, we do not want to assume that there exists a representative agent of these different individuals as is generally assumed when using the Ramsey equation. This raises the issue of the inequality within future generations. One standard methodology to take inequality into account relies on *equity weights* when we compute the future costs and benefits of a policy (see Anthoff, Hepbrun and Tol 2009 for an application to climate policy). In that case, we compute an "average" social discount rate for the future (that looks at the average level of a good enjoyed by the future generation) and transform the measure of impacts in the future to account for inequalities. An alternative consist in incorporating directly (intra-generational) inequality considerations in the measure of the social discount rate as suggested by Gollier (2015) and Fleurbaey and Zuber (2015a). The idea is that a less unequal future generation can be considered as better-off so that we may want to put less weight in increases in their consumption of goods and resources. In that case, the social discount rate can be viewed as an aggregation of individual or personalized discount rates (for person-to-person

transfers) and it also makes it possible to include the consideration of inequalities in the distribution of consequences (Fleurbaey and Zuber 2015a).

Risk is also a pervasive phenomenon in many problems involving intergenerational ethics and in particular for climate justice. There is a lot of uncertainty about the level of resources available to future uncertainty and also epistemic uncertainty about the models we can use to foresee the future consequences of our actions. Risk has attracted a lot of attention in the economic literature. Several attempts have been made to adjust the discounting formula in that case as well as providing alternative decision model that may disentangle attitudes towards risk and attitudes towards intergenerational distribution in the value function (see Greaves 2017 and Fleurbaey et al. 2019 for extensive surveys of this issue). Formally, the question of risk is very similar to the question of inequality. But there is one key difference: in the case of risk, the discounting formula may include an additional term that reflects the correlation between individual advantage and the aggregate advantage of people in all generations (Fleurbaey and Zuber 2015b). This involves a notion of correlated or aggregate risk, which is a risk on the overall value function.

3 Conclusion: Beyond social discounting

In this chapter, we have argued that social discounting is a flexible tool for policy evaluation for problems spanning several generations. Social discounting does not necessarily entail a violation of the principle of impartiality among generations: on the contrary, we have argued that ethically defensible versions of social discounting typically satisfy this principle. Social discounting may also involve different ethical positions regarding how resources, well-being or advantages should be distributed across people in different generations. We have actually argued that the social discounting methodology can be developed for a wide range of ethical views. The main restriction is that it remains within the scope of consequentialist axiology, and thus is not the appropriate tool to deal with considerations of right, harm, virtues or duties. But even people who do not view consequentialism as the only relevant moral theory may accept that consequentialist considerations are part of the overall moral picture: in that sense, the social discounting methodology may still be useful for them.

However, even within the scope of cost-benefit analysis, the social discounting approach (and the associated methodology of net present values) should not be considered an all-purpose tool that can serve for all evaluations and issues. We would like to conclude this chapter with some caveats.

First, we must emphasize a well-known point that is sometimes overlooked. As stressed for instance by the Stern review (Stern 2006 *§* 2A.2), the social discount

rate is useful to evaluate small transfers of consumption across individuals living at different times. It is not adapted to large-scale changes. For instance, there is evidence that climate change may affect future growth and therefore that a climate policy may alter the underlying consumption path (Dell et al. 2012, Moore and Diaz 2015). Similarly, climate change may worsen inequalities or hinter the development of some regions in the world, and such changes may alter our willingness to implement policies that reduce more greenhouse gas emissions in the near term (Hallegate et al. 2016, Budolfson et al. forth.).

Similarly, policy may change the size or the composition of the future population. For instance, climate change and climate policy could influence patterns of fertility and mortality thereby changing who will exist in the future. Thus, our value function should incorporate population sizes, thereby raising issues of population ethics (Broome 2012, Kolstad et al. 2014, p. 211). Population ethics is known to raise difficult puzzles and no single approach has emerged that is consistent with all attractive intuitions (Parfit 1984; Blackorby et al. 2005). A broad divide is between theories that value population size even at the expense of average well-being (like Total Utilitarianism) and theories that regard average well-being as the most important aspect even if it implies reducing population size (like Average Utilitarianism). Population ethics can significantly modify our view on policy especially in cases when we are not sure about the future population trajectory (Scovronick et al. 2017, Méjean et al. 2017).

Population ethics is particularly important in the case of catastrophic or existential risks that may drastically reduce future population size (or even lead to human extinction). Climate policy in general has focused on future impacts of climate on consumption or on the goods (in a broad sense) available to future generations. But climate change may not only alter future resources, it may alter the risk that future generations (not only human, but also for other species) do not exist. In economic cost-benefit analysis, the technique used to evaluate changes in probabilities of a risk on the existence (of an individual) consists in computing the 'value of a statistical life'. The methodology can be extended in the case of risks on the existence of future generations (the idea was suggested, but used in a very different way by Weitzman, 2009). In that case, the social discount rate is not any more the key parameter to value future consequences and alternative methods must be developed (Bommier, Lanz and Zuber 2015; Méjean et al. 2017).

For all three cases mentioned above (effects of the policy on growth, on population size or on large-scale risk), the computation of net present values using some social discount rate cannot provide the right guidance if the underlying ethical theory can be represented by a value function like the one exhibited in Formulas (1) or (4). For these cases (and other similar cases) policy evaluation has to rely directly

on the underlying value function. This reminds us that the social discounting methodology is only an approximation of our ethical assessment. And that the soundness or attractiveness of a discounting formula only derives from the soundness and attractiveness of the underlying ethical theory. Discounting is good only insofar as it relies on sound ethical principles.[15]

References

Adler, M., D. Anthoff, V. Bosetti, G. Garner, K. Keller, and N. Treich (2017). Priority for the Worse Off and the Social Cost of Carbon, Nature Climate Change 7: 443–449.

Adler, M. and M. Fleurbaey (2016). The Oxford Handbook of Well-Being and Public Policy. Oxford: Oxford University Press.

Alkire, S. (2016). The capability approach and well-being measurement for public policy. In M. Adler and M. Fleurbaey (Eds), The Oxford Handbook of Well-Being and Public Policy, Oxford: Oxford University Press.

Anthoff, D., C. Hepburn and R.S.J. Tol, (2009). Equity weighting and the marginal damage costs of climate change. Ecological Economics 68 : 836–849.

Arrow, K.J. (1999). Discounting, morality and gaming. In P.R. Portney and J.P. Weyant (Eds.), Discounting and Intergenerational Equity, New York: Resources for the Future.

Arrow, K.J., W.R. Cline, K.-G. Mäler, M. Munasinghe, R. Squitieri and J.E. Stiglitz (1996). Intertemporal equity, discounting, and economic efficiency. In J.P. Bruce, H. Lee and E.F. Haites (Eds.), Climate Change 1995. Economic and Social Dimensions of Climate Change. Contribution of Working Group III to the Second Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge: Cambridge University Press.

Asheim, G.B. and W. Buchholz (2003). The malleability of undiscounted utilitarianism as a criterion of intergenerational justice. Economica 70: 405–422.

Asheim, G.B., T. Mitra and B. Tungodden (2012). Sustainable recursive social welfare functions. Economic Theory 49: 267–292.

[15] Acknowledgements: This research has been supported by the Agence nationale de la recherche through the Fair-ClimPop project (ANR-16-CE03-0001-01) and the Investissements d'Avenir program (ANR-17-EURE-01). Financial support by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences) is also gratefully acknowledged.

Atkinson, G., S. Dietz, J. Helgeson, C. Hepburn and H. Saelen (2009). Siblings, not triplets: Social preferences for risk, inequality and time in discounting climate change. Economics: The Open-Access, Open-Assessment E-Journal 3(2009-6).

Basu, K. and T. Mitra (2003). Aggregating infinite utility streams with intergenerational equity: The impossibility of being Paretian. Econometrica 71: 1557–1563.

Blackorby, C., W. Bossert and D. Donaldson (2005). Population Issues in Social Choice Theory, Welfare Economics, and Ethics. Cambridge: Cambridge University Press.

Bommier, A., B. Lanz and S. Zuber (2015). Models-as-usual for unusual risks? On the value of catastrophic climate change. Journal of Environmental Economics and Management 74:1–22.

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

Broome, J. (2012). Climate Matters: Ethics in a Warming World. New York: Norton.

Broome, J. (2015). Equality versus priority: a useful distinction. Economics and Philosophy 31: 219–228.

Budolfson, M., F. Dennig, M. Fleurbaey, N. Scovronick, A. Siebert., D. Spears and F. Wagner (forthcoming). Optimal climate policy and the future of world economic development. World Bank Economic Review.

Chichilnisky, G. (1996). An axiomatic approach to sustainable development. Social Choice and Welfare 13: 231–257.

Crisp, R. (2003). Equality, Priority, and Compassion. Ethics 113: 745–763.

Dasgupta, P. (2008). Discounting Climate Change. Journal of Risk and Uncertainty 37:141–169.

Dasgupta, P. (2016). Birth and Death. Unpublished manuscript.

Dasgupta, P. and G. Heal (1979). Economic theory and exhaustible resources. Cambridge: Cambridge University Press

Dell, M., B. Jones and B Olken (2012). Temperature shocks and economic growth: evidence from the last half century. American Economic Journal: Macroeconomics 4(3): 66–95.

Diamond, P. (1965). The evaluation of infinite utility streams. Econometrica 33: 170–177.

Dietz, S. and G.B. Asheim (2012). Climate policy under sustainable discounted utilitarianism. Journal of Environmental Economics and Management, 63(3): 321–335.

Drummond, M.F., G.L. Stoddart and G.W. Torrance (1987). Methods for the Economics Evaluation of Health Care Programmes. Oxford: Oxford University Press.

Fleurbaey, M. and D. Blanchet (2013). Beyond GDP. Measuring Welfare and Assessing Sustainability. Oxford: Oxford University Press.

Fleurbaey, M., Ferranna M., Budolfson M., Dennig F., Mintz-Woo K., Socolow R., Spears D., Zuber S. (2019). The social cost of carbon: Valuing inequality, risk, and population for climate policy. The Monist 102: 84–109.

Fleurbaey, M. and F. Maniquet (2011). A Theory of Fairness and Social Welfare. Cambridge: Cambridge University Press.

Fleurbaey, M. and P. Mongin (2016). The utilitarian relevance of Harsanyi's theorem. AEJ: Microeconomics 8: 289–306.

Fleurbaey, M. and S. Zuber (2013). Climate policies deserve a negative discount rate. Chicago Journal of International Law 13: 565–595.

Fleurbaey, M. and S. Zuber (2015a). Discounting, risk and inequality: A general approach. Journal of Public Economics 128: 34–49.

Fleurbaey, M. and S. Zuber (2015b). Discounting beyond utilitarianism. Economics: The Open-Access, Open-Assessment E-Journal 9 (2015-12): 1–52.

Frankfurt, H. (1987). Equality as a Moral Ideal. Ethics 98: 21–42.

Frankfurt, H. (2000). The Moral Irrelevance of Equality. Public Affairs Quarterly 14: 87–103.

Gollier, C. (2010). Ecological discounting. Journal of Economic Theory 145: 812–829.

Gollier, C. (2015). Discounting, inequality and economic convergence. Journal of Environmental Economics and Management 69: 53–61.

Greaves, H. (2017). Discounting for public policy: A survey. Economics and Philosophy 33: 391–339.

Hallegatte, S., M. Bangalore, L. Bonzanigo, M. Fay, T. Kane, U. Narloch, J. Rozenberg, D. Treguer and A. Vogt-Schilb (2016). Shock Waves: Managing the Impacts of Climate Change on Poverty. Climate Change and Development Series, Washington, DC: World Bank.

Harsany, J. (1953). Cardinal utility in welfare economics and in the theory of risk-taking. Journal of Political Economy 61: 434–435.

Harsanyi, J. (1955). Cardinal welfare, individualistic ethics and interpersonal comparison of utilities. Journal of Political Economy 63: 309–321.

Hsiung, W. and C. Sunstein (2007). Climate change and animals. University of Pennsylvania Law Review 155: 1695–1740.

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. In O. Edenhofer, 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.), Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press.

Koopmans, T.J. (1960). Stationary ordinal utility and impatience. Econometrica 28: 287–309.

Lauwers, L. (2010). Ordering infinite utility streams comes at the cost of a non-Ramsey set. Journal of Mathematical Economics 46: 32–37.

McShane, K. (forthcoming). Why animal welfare is not biodiversity, ecosystem services, or human welfare: toward a more complete assessment of climate impacts. Les Ateliers de l'Ethique/The Ethics Forum.

Méjean, A., A. Pottier, M. Fleurbaey and S. Zuber (2017). Intergenerational equity under catastrophic climate change. CES working paper #2017.40.

Mirrlees, J.-A. (1967). Optimum growth when technology is changing. Review of Economic Studies 34: 95–124.

Millner, A. (2013). On welfare frameworks and catastrophic climate risks. Journal of Environmental Economics and Management, 65: 310–325.

Moore, F. and D. Diaz (2015). Temperature impacts on economic growth warrant stringent mitigation policy, Nature Climate Change 5: 127–131.

Nagel, T. (2005). The problem of global justice. Philosophy and Public Affairs 33: 113–147.

Nordhaus, W.D. (2007). A review of the Stern Review on the Economics of Climate Change. Journal of Economic Literature 45: 686–702.

Nordhaus, W.D. (2008). A Question of Balance: Weighing the Options on Global Warming Policies. New Haven: Yale University Press.

Okun, A.M. (1975). Equality and Efficiency: The Big Trade-Off. Washington (DC): Brookings Institution.

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

Parfit, D. (1997). Equality or priority? Ratio, 10, 202–221.

Posner, E.A. and D. Weisbach (2000). Climate Change Justice. Princeton: Princeton University Press.

Ramsey, F.P. (1928). A mathematical theory of savings. Economic Journal 38: 543–559.

Rawls, J. (1971). A Theory of Justice, Revised Edition (1999). Cambridge: The Belknap Press of the Harvard University Press.

Rawls, J. (1999a). The Law of Peoples. Cambridge: Harvard University Press.

Roemer, J. (1996). Theories of Distributive Justice. Cambridge: Harvard University Press.

Schelling, T. (1995). Intergenerational discounting. Energy Policy 23: 395–401.

Scovronick, N., M.B. Budolfson, F. Dennig, M. Fleurbaey, A. Siebert, R.H. Socolow, D. Spears and F. Wagner (2017). Impact of population growth and population ethics on climate change mitigation policy. Proceedings of the National Academy of Sciences 114(46): 12338–12343.

Sebo, J. (forthcoming). Animals and Climate Change. In M. Budolfson, T. McPherson and D. Plunkett (Eds.), Philosophy and Climate Change. Oxford: Oxford University Press.

Sen, A.K. (1976). Welfare inequality and Rawlsian axiomatics. Theory and Decision 7: 243–262.

Sen, A.K. (1985). Commodities and Capabilities. Amsterdam: North Holland.

Sidgwick H. (1907). The Methods of Ethics, Seventh Edition. London: MacMillan.

Stern, N. (2006). Stern Review: The Economics of Climate Change, Volume 30. London: Her Majesty's Treasury.

Temkin, L.S. (1993). *Inequality*. Oxford: Oxford University Press.

Tol, R.S.J. (2010). International inequity aversion and the social cost of carbon. Climate Change Economics 1: 21–32.

Von Weizsäcker C.C. (1965). Existence of optimal programs of accumulation for an infinite time horizon. Review of Economic Studies 32: 85–104.

Weitzman, M. (2009). On modeling and interpreting the economics of catastrophic climate change. The Review of Economics and Statistics 91: 1–19.

Zame, W.R. (2007). Can intergenerational equity be operationalized? Theoretical Economics 2: 187–202.

Zuber, S. (2011). The aggregation of preferences: Can we ignore the past? Theory and Decision 70: 367–384.

Zuber, S. and G.B. Asheim (2012). Justifying social discounting: The rank-discounted utilitarian approach. Journal of Economic Theory 147: 1572–1601.

Stéphane Zuber[1]

Population-Adjusted Egalitarianism[2]

Egalitarianism focuses on the well-being of the worst-off person. It has attracted a lot of attention in economic theory, for instance when dealing with the sustainable intertemporal allocation of resources. Economic theory has formalized egalitarianism through the Maximin and Leximin criteria, but it is not clear how they should be applied when population size may vary. In this paper, I present possible justifications of egalitarianism when considering populations with variable sizes. I then propose new versions of egalitarianism that encompass many views on how to trade-off population size and well-being. I discuss some implications of egalitarianism for optimal population size. I first describe how population ethical views affects population growth. In a model with natural resources, I then show that utilitarianism always recommend a larger population for low levels of resources, but that this conclusion may not hold true for larger levels.

[1] Paris School of Economics – CNRS, France. E-mail: Stephane.Zuber@univ-paris1.fr.

[2] This research has been supported by the Agence nationale de la recherche through the Fair-ClimPop project (ANR-16-CE03-0001-01), Investissements d'Avenir program (ANR-10-LABX-93). Financial support from the Swedish foundation for humanities and social sciences is also gratefully acknowledged (Anslag har erhållits från Stiftelsen Riksbankens Jubileumsfond). I would like to thank seminar and conference audience at Institute for Advance Studies (Marseille), Centre d'Economie de la Sorbonne (Paris), Meeting of Society for Social Choice and Welfare (Seoul) and Institute for Futures Studies (Stockohlm) for their comments.

1 Introduction

Egalitarianism is an important principle of social justice that promotes an equal (and efficient) distribution of resources. It has attracted a lot of attention in contemporary moral philosophy since Rawls (1971), even if there have been discussions about what exactly should be equalized (Sen, 1980). In economic theory, egalitarianism has been modeled either through a Maximin criterion or through a lexicographic version of Maximin named Leximin. Many different axiomatic characterizations of such egalitarian criteria can be found in the literature (see for instance Hammond, 1976, Sen, 1986, Barberá and Jackson, 1988, Lauwers, 1997, D'Aspremont and Gevers, 2002, Fleurbaey and Maniquet or 2011).

Egalitarian criteria have been considered by economic theory to deal with the optimal allocations of resources, in particular in an intergenerational context where sustainability issues may arise. Solow (1975) characterized egalitarian intergenerational distributions in a model with an exhaustible resource and showed that they lead sustainable (actually constant) levels of consumption in contrast to utilitarian solutions. The Maximin path, if egalitarian and efficient, indeed satisfies Hartwick's sustainability rule, which requires investing rents from exhaustible resources in reproducible capital to compensate for the depletion of their stocks (Hartwick, 1977).[1]

A key question for sustainable development and the intertemporal allocation of resources is however population size. In particular, concerns about global climate change have renewed the interest in assessing the impacts of policy on population (see for instance IPCC, 2015a, chap. 11) and in the normative aspects of population size (see for instance IPCC, 2015b, chap. 3) The problem then for the egalitarian perspective is to define how the Maximin or Leximin should be applied when population size may vary. There exist few attempts to define such egalitarian rules in a variable population context (Bossert, 1990; Blackorby, Bossert and Donaldson, 1996). However, existing criteria have serious drawbacks (Blackorby, Bossert and Donaldson, 2005; Arrhenius, forth.). According to Critical-level Leximin, as defined by Blackorby, Bossert and Donaldson (1996), any population with excellent lives is worse than a population with one additional person even when the well-being of all the individuals in the latter population is barely above a critical level. According to the Maximin proposed by Bossert (1990) (and the corresponding Leximin suggested by Arrhenius, forth., Sect. 6.8), any population is worse than a population consisting of one individual, provided that the worst-off individual of the former has lower well-being than the single individual of the latter.

In this paper, I propose new versions of egalitarianism that encompass many views about how to trade-off population size and well-being. First, in Section 2, I present arguments to justify egalitarianism when considering populations with variable sizes. One line of argument is similar to that of Fleurbaey and Tungodden (2010): if we satisfy a minimal non-aggregation property that limits the loss by the worst-off for the sake of all best-offs, we are compelled to egalitarian criteria under a consistency requirement. Another (new) line of argument is that, if we accept that the the best-off should make limited sacrifice for the sake of a sufficiently large number of worst-offs, we are also compelled to egalitarian criteria under the consistency requirement.

In Section 3, I discuss how to compare populations with different sizes, provided

[1] See also, for general proofs, Withagen and Asheim (1998) or Mitra (2002).

we use a Maximin criterion. Blackorby, Bossert and Donaldson (1996) have proposed critical-level properties to compare populations with different sizes. I follow this route, and use a weak critical-level property together with a condition on utility measurement to describe a large new class of Maximin social welfare orderings avoiding the repugnant conclusion described by Parfit (1984). The idea is to multiply individuals' well-being by a weight that depends on population size (provided well-being if non-negative) and then to apply a Maximin like in Bossert (1990). Doing so, I am able to cover a variety of attitudes towards the trade-off between population size and (minimal) well-being, avoiding the problems of previous criteria.

In Section 4, I study some implications of egalitarian social welfare orderings for optimal population size. I first provide a general condition on population ethics views (embodied in a function aggregating population size and an equally-distributed equivalent welfare measure) that guarantees that we can order social welfare functions of the same class in terms of optimal population size, whatever the specific underlying economic model of resource allocation. I then compare egalitarian and utilitarian criteria in a simple model with a renewable resource. I show that utilitarian criteria always recommend a larger population than egalitarian criteria for a specific population ethics view or when the level of resources is low. However, a numerical example shows that this finding is not true in general. It is not possible to say that utilitarianism always entails larger population sizes.

Section 5 concludes. The proofs of the main results are in Appendix A. Supplementary materials contain additional results, in particular a proof of the independence of the axioms in Theorem 1 and an analysis of a Leximin counterpart of the Maximin criteria discussed in Section 3.

2 Justifying egalitarianism when population size may vary

Let $\mathbb{N}$ denote the set of positive integers and $\mathbb{R}$ (resp. $\mathbb{R}_+$, $\mathbb{R}_{++}$, $\mathbb{R}_-$, $\mathbb{R}_{--}$) denote the set of real numbers (resp. non-negative, positive, non-positive, negative real numbers). I also let $I_n = \{1, \cdots, n\}$. Let $X = \cup_{n\in\mathbb{N}}\mathbb{R}^n$ be the set of possible finite *allocations* of lifetime well-being. For every $n \in \mathbb{N}$, each allocation $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$ is associated with a finite population size, $n(x) = n$. Following the usual convention in population ethics, a lifetime well-being level equal to 0 represents *neutrality*. Hence, lifetime well-being is normalized so that above neutrality, a life, as a whole, is worth living; below neutrality, it is not. I also define X_+ as the set of allocations such that all individuals have non-negative well-being levels, i.e. $X_+ = \{x \in X | x_i \geq 0, \forall i \in I_{n(x)}\}$. Similarly, let $X_- = \{x \in X | x_i \leq 0, \forall i \in I_{n(x)}\}$.

A social welfare ordering (henceforth SWO) on the set X is a complete, reflexive and transitive binary relation $\succsim$, where for all $x, y \in X$, $x \succsim y$ means that the allocation x is deemed socially at least as good as y. Let $\sim$ and $\succ$ denote the symmetric and asymmetric parts of $\succsim$.

For each $x \in X$, $x_{[\,]} = (x_{[1]}, \ldots, x_{[r]}, \ldots, x_{[n(x)]})$ denotes the non-decreasing allocation, which reorders the components of x; i.e., for each rank $r \in I_{n(x)-1}$, $x_{[r]} \leq x_{[r+1]}$. For every $n \in \mathbb{N}$ and all $x, y \in \mathbb{R}^n$, we write $x_{[\,]} \geq y_{[\,]}$ whenever $x_{[r]} \geq y_{[r]}$ for all $r \in I_{N(x)}$;

we write $x_{[\,]} > y_{[\,]}$ whenever $x_{[\,]} \geq y_{[\,]}$ and $x_{[\,]} \neq y_{[\,]}$; and we write $x_{[\,]} \gg y_{[\,]}$ whenever $x_{[r]} > y_{[r]}$ for all $r \in I_{N(x)}$.

For any $a \in \mathbb{R}$ and $n \in \mathbb{N}$, $(a)_n$ denotes the allocation $(a, \cdots, a) \in \mathbb{R}^n$. For any $\lambda \in \mathbb{R}_{++}$ and $x \in X$, let denote λx the allocation y such that $n(y) = n(x)$ and $y_i = \lambda x_i$ for all $i \in I_{n(x)}$. For any $x, y \in X$, (x, y) denotes an allocation $z \in X$ such that $n(z) = n(x) + n(y)$, $z_i = x_i$ for all $i \in I_n(x)$ and $z_i = y_{i-n(x)}$ for all $i \in I_n(z) \setminus I_n(x)$. Hence (x, y) corresponds to a situation where a population with allocation y is added to an existing population with allocation x. In particular $\big(x, (a)_1\big)$ corresponds to a situation where a single person with well-being $a \in \mathbb{R}$ is added the existing population with allocation x.

Let us now introduce the definitions of the two egalitarian social welfare orderings for fixed populations, namely the Maximin and Leximin social welfare orderings.

Definition 1 *For $n \in \mathbb{N}$, the Maximin SWO on $\mathbb{R}^n$, denoted $\succsim_M^n$, is defined as follows. For all $x, y \in \mathbb{R}^n$, $x \succsim_M^n y$ if and only if $x_{[1]} \geq y_{[1]}$.*

We say more generally that an SWO $\succsim$ on X is a Maximin SWO *if for all $n \in \mathbb{N}$ and for all $x, y \in \mathbb{R}^n$, $x \succsim y$ if and only if $x \succsim_M^n y$.*

Definition 2 *For $n \in \mathbb{N}$, the Leximin SWO on $\mathbb{R}^n$, denoted $\succsim_L^n$, is defined as follows. For all $x, y \in \mathbb{R}^n$,*

(a) *$x \sim_L^n y$ if and only if $\big(x_{[1]}, \ldots, x_{[n]}\big) = \big(y_{[1]}, \ldots, y_{[n]}\big)$.*

(b) *$x \succ_L^n y$ if and only if there exists $R \in I_n$ such that $x_{[r]} = y_{[r]}$ for all $r \in I_{R-1}$ and $x_{[R]} > y_{[R]}$.*

We say more generally that an SWO $\succsim$ on X is a Leximin SWO *if for all $n \in \mathbb{N}$ and for all $x, y \in \mathbb{R}^n$, $x \succsim y$ if and only if $x \succsim_L^n y$.*

These egalitarian social welfare orderings have been justified by Hammond (1976) with the following principle.

Hammond Equity. For all $n \in \mathbb{N}$, for all $x, y \in \mathbb{R}^n$, if $y_i < x_i < x_j < y_j$ for some $i, j \in I_n$ and $x_k = y_k$ for all $I_n \setminus \{i, j\}$ then $x \succ y$.

Hammond Equity is a strong equity requirement, that allows large losses in total utility for the sake of well-being equalization. It is thus often considered too extreme. To obtain justifications of egalitarianism, I will consider equity requirements, stating that limited sacrifice of the best-off(s) are acceptable provided that the gains by the worst-off(s) are sufficient, either because the single worst-off benefit sufficiently or because there are sufficiently many worst-offs benefiting. The first of my equity principle actually also aims at protecting current generations against unlimited sacrifices for the sake of future generations. It states that, whenever the current generation is worse-off, there is a bound on the loss the society can require for a sufficient gain experienced by all future generations.

Limited sacrifice for the rich future. For all $\alpha \in \mathbb{R}{+}{+}$ there exist $\alpha > \beta > 0$ such that, for all $n \in \mathbb{N}$, if $a, b, c, d \in \mathbb{R}$ are such that $b \leq c$, $b - a \geq \alpha$ and $\beta \geq d - c$, then $\big((b)_1, (c)_n\big) \succsim \big((a)_1, (d)_n\big)$.

Limited sacrifice for the rich future is related to the principle of Mild non-aggregation discussed by Fleurbaey and Tungodden (2010). It simplifies their formulation by considering only two classes of people: a single worst-off and the rest of the population, which is equally well-off. Limited sacrifice for the rich future is also a weakening of Hammond Equity because it imposes bounds on how much the best-offs sacrifice for the sake of the worst-off, and how much the worst-off gains.[2]

Limited sacrifice for the long future. There exists $\gamma \in \mathbb{R}_{++}$ and $k \in \mathbb{N}$ such that, for all $n \in \mathbb{N}$ and $a, b, c, d \in \mathbb{R}$, if $a < b \leq c, n \geq k$ and $d - c \leq \gamma$, then $\big((c)_1, (b)_n\big) \succsim \big((d)_1, (a)_n\big)$.

Limited sacrifice for the long future means that if the cost for the best-off is limited (less than γ) and if the number of poor who gain is sufficiently large, we always want to make the transfer (even though the poor may not gain much). It is comparable to the axiom of Hammond Equity for Future introduced by Asheim, Mitra and Tungodden (2007) in the context of the evaluation of infinite utility streams. Hammond Equity for Future states that, if the present is better-off than the future and a sacrifice now improve the well-being of all future generations while leaving the present generation relatively better-off, then such a transfer is socially desirable. The difference with Limited sacrifice for the rich future is that we now have a finite (but large) number of generations; but we limit the sacrifice made by the best-off generation.

To obtain a justification for egalitarianism, I also assume that we endorse the following two principles.

Suppes-Sen. For all $n \in \mathbb{N}$ and $x, y \in \mathbb{R}^n$, if $x_{[]} \geq y_{[]}$, then $x \succsim y$; if $x_{[]} \gg y_{[]}$, then $x \succ y$.

The Suppes-Sen principle represents two ideas. First, that individual and generations should be treated in the same (anonymous) way, so that permuting welfare levels has no impact on the social evaluation. Second, that situations where all individuals have a higher level of welfare can never be worse.

The second principle is a principle of consistency across populations: whenever a population can be split in two subpopulations and that both subpopulations are at least as well-off in one alternative than in another, then the first alternative is weakly better. If both subpopulations are strictly better-off, so is the aggregate population.

Consistency. For all $n, m \in \mathbb{N}$, all $x, y \in \mathbb{R}^n$ and all $x', y' \in \mathbb{R}^n$, if $x \succsim y$ and $x' \succsim y'$ then $(x, x') \succsim (y, y')$. If furthermore $x \succ y$ and $x' \succ y'$ then $(x, x') \succ (y, y')$.

I then obtain a first egalitarian result.

Proposition 1 *Consider an SWO* $\succsim$ *on* X.

[2]Limited sacrifice for the rich future can still be considered strong as any level of sacrifice of the current generation is possible provided the future gain is large enough. However, I show in the Supplementary material (Section S.A) that this principle is implied by a weaker principle related to the Weak non-aggregation principle of Fleurbaey and Tungodden (2010) provided we accept a principle of Ratio-scale invariance. Given that I will use Ratio-scale invariance for variable population comparisons, it seems natural to use the stronger Limited sacrifice for the rich future principle at this stage.

1. *If $\succsim$ satisfies Suppes-Sen, Consistency and Limited sacrifice for the rich future, then, for all $n \in \mathbb{N}$ and $x, y \in \mathbb{R}^n$, $x \succ y$ whenever $x_{[1]} > y_{[1]}$.*

2. *If $\succsim$ satisfies Suppes-Sen, Consistency and Limited sacrifice for the long future, then, for all $n \in \mathbb{N}$ and $x, y \in \mathbb{R}^n$, $x \succ y$ whenever $x_{[1]} > y_{[1]}$.*

To obtain a complete characterization of the maximin social welfare ordering, one only needs to add a continuity requirement.

Continuity. For any $n \in \mathbb{N}$ and any $x \in \mathbb{R}^n$, the sets $\{y \in \mathbb{R}^n | x \succsim y\}$ and $\{y \in \mathbb{R}^n | y \succsim x\}$ are closed.

Theorem 1 *Consider an SWO $\succsim$ on X.*

1. *$\succsim$ satisfies Suppes-Sen, Consistency, Continuity and Limited sacrifice for the rich future, if and only if it is a Maximin SWO.*

2. *$\succsim$ satisfies Suppes-Sen, Consistency, Continuity and Limited sacrifice for the long future, if and only if it is a Maximin SWO.*

Appendix B shows that the two characterizations above are tight in the sense that all principles involved in the characterization are independent from one another.

3 Egalitarianism for variable populations

For the moment, we have only compared populations with the same size. In this section, I introduce and characterize Population-adjusted maximin SWOs that can be used to assess allocations when population size varies. First, let us define a function that transforms well-being numbers according to population size:

> Let $\kappa = (\kappa_n)_{n \in \mathbb{N}}$ be a sequence of real numbers. Π_κ denotes the function $\Pi_\kappa : \mathbb{N} \times \mathbb{R}$ such that, for all $(n, e) \in X \times \mathbb{R}$:

$$\Pi_\kappa(n, e) = \begin{cases} e & \text{if } e \leq 0; \\ \kappa_n \cdot e & \text{if } e > 0. \end{cases}$$

Definition 3 *An SWO $\succsim$ on X is a Population-adjusted maximin SWO if and only if there exists a non-decreasing sequence $(\kappa_k)_{k \in \mathbb{N}}$ such that $\kappa_1 = 1$ and for all $x, y \in X$,*

$$x \succsim y \iff \Pi_\kappa\big(n(x), x_{[1]}\big) \geq \Pi_\kappa\big(n(y), y_{[1]}\big).$$

Population-adjusted maximin SWOs use the usual maximin procedure by comparing the minimal level well-being in allocations that may have different sizes. However, the contributive value of a life is adjusted by population size before comparing these minimal well-being levels. The adjustment procedure is very simple. If well-being is negative, no adjustment is made. If well-being is positive, it is multiplied by a factor that depends on population size (with larger populations having a larger weight).

To characterize these Population-adjusted maximin SWOs, I introduce two additional principles. A first natural principle is based on the notion of a critical level, that is a level such that adding an individual at that level is a matter of social indifference. The following principle states that such a critical level always exists (but may vary depending on population size and the existing allocation).

Critical level. For any $x \in X$ there exists $a \in \mathbb{R}_+$ such that $\left(x, (a)_1\right) \sim x$.

Another principle is that changes in the scale of the measurement of individual well-being do not affect the social ranking.

Ratio-scale invariance. For any $\lambda \in \mathbb{R}_{++}$ and any $x, y \in X$, $x \succsim y$ if and only if $\lambda x \succsim \lambda y$.

Ratio-scale invariance is a property about the measurement of individual well-being. It implies that only the ratios of well-being levels are directly fully comparable. It also involves picking an interpersonally significant norm, which is the 0. This makes sense in the context of population ethics, where 0 level is a well-being level at which life is no more worthwhile than death, and is supposed to be normatively comparable across people. Note that a similar property has often been consider in the literature with a fixed population (see, e.g. Roberts , 1980; Blackorby and Donaldson, 1982)

Theorem 2 *A Maximin SWO $\succsim$ on X satisfies Critical level and Ratio-scale invariance if and only if it is a Population-adjusted maximin SWO.*

Again, Appendix B shows that the characterization is tight (and more precisely that characterizations using principles in Th. 1 and Th. 2 are tight).

It is interesting to discuss the population ethics properties of the Population-adjusted maximin SWOs characterized in Th. 2. The literature on population ethics has indeed highlighted that variable population social welfare criteria may face several ethical issues. The most discussed issue is the *repugnant conclusion*. According to Parfit (1984), a social welfare ordering leads to the repugnant conclusion if:

> "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."

The repugnant conclusion has caught much attention in the literature on population ethics as most of the literature have discussed ways to avoid such a conclusion. Formally, one can formulate avoidance of the repugnant conclusion as follows:[3]

Avoidance of the repugnant conclusion. An SWO $\succsim$ avoids the repugnant conclusion if there exist $k \in \mathbb{N}$, $a \in \mathbb{R}_{++}$ and $b \in [0, a]$ such that, for all $m \geq k$, $(a)_k \succsim (b)_m$.

Another problem that an SWO may face is that it may imply the *sadistic conclusion*. According to Arrhenius (2000), an SWO leads to the sadistic conclusion in the following case:

[3] Other formalizations have been proposed, for instance by Blackorby, Bossert and Donaldson (2005). The formulation is slightly stronger than the one they have, but I think it is in the spirit of Parfit's initial formulation.

> "When adding people without affecting the original people's welfare, it can be better to add people with negative welfare rather than positive welfare."

Formally, one can formulate avoidance of the sadistic conclusion as follows:[4]

Avoidance of the sadistic conclusion. An SWO $\succsim$ avoids the sadistic conclusion if for all $x \in X$, all $k, m \in \mathbb{N}$, all $a \in \mathbb{R}_{++}$ and $b \in \mathbb{R}_{--}$, $\big(x, (a)_k\big) \succsim \big(x, (b)_m\big)$.

Lastly, we may want to have principles about how the addition of a person changes social welfare depending on whether her well-being level if positive or negative. First, it may seem natural that adding someone with positive well-being to a population without affecting existing people well-being is always acceptable: no one else is affected and the additional person has a good enough life. This is known as the *Mere addition principle*. Second, it seems natural on the contrary that we do not want to add someone with a negative level of well-being to a population: this person has a life not worth living. Arrhenius (forth.) named this principle the *Negative mere addition principle* (it is called the expansion principle by Blackorby, Bossert and Donaldson, 2005) The formal statements of these principles are as follows:

Mere addition principle. An SWO $\succsim$ satisfies the Mere addition principle if for all $x \in X$ and all $a \in \mathbb{R}_{++}$, $\big(x, (a)_1\big) \succsim x$.

Negative mere addition principle. An SWO $\succsim$ satisfies the Negative mere addition principle if for all $x \in X$ and all $a \in \mathbb{R}_{--}$, $x \succ \big(x, (a)_1\big)$.

It turns out that Population-adjusted maximin SWOs can avoid both the sadistic conclusion and the repugnant conclusion. On the other hand, they necessarily violate the Mere addition principle and the Negative mere addition principle.

Proposition 2 *Assume that the SWO $\succsim$ is a Population-adjusted maximin SWO. Then it always avoids the sadistic conclusion and it also avoids the repugnant conclusion if and only if the sequence $(\kappa_k)_{k\in\mathbb{N}}$ in Def. 3 is bounded. However, $\succsim$ satisfies neither the Mere addition principle nor the Negative mere addition principle.*

Violating the Negative mere addition principle is probably not a good feature of Population-adjusted maximin SWOs. In the supplementary material (Section S.B), I introduce and characterize Population-adjusted versions of Leximin that satisfy the Negative mere addition principle.

The appeal of the Mere addition principle has been discussed in the literature. In particular, Carlson (1998) proved that the Mere addition principle and a Non-anti egalitarianism principle imply a conclusion akin to the Repugnant conclusion. Indeed adding a person with a low level of well-being to an otherwise well-off population may increase inequality, which may not be an improvement from the social point of view. Note that Population-adjusted maximin SWOs satisfy the following weak version of the Mere addition principle that holds when people in the existing population only have negative levels of well-being.

Weak mere addition principle. An SWO $\succsim$ satisfies the Weak mere addition principle if for all $x \in X_-$ and all $a \in \mathbb{R}_{++}$, $\big(x, (a)_1\big) \succsim x$.

[4] Note that we interpret "better" in the statement of the sadistic conclusion as "strictly better".

4 Egalitarianism, optimal population size and natural resources

Optimal population size: a simple model

The question of optimal population has attracted a lot of attention in the economic literature. Already Malthus hypothesized that natural population growth (which is supposedly geometric) is constrained by economic growth (which is only arithmetic) yielding recurrent episodes of extended poverty and migration (Malthus, 1798). Wicksell shared this view and advocated to limit natality through the systematic use of contraceptives within the marriage. The objective was to reach an optimum population size that Wicksell (like many economists) conceived as the one maximizing average well-being (Wicksell, 1893).[5] Later on, Meade (1955) discussed another criteria of optimum population size, namely the maximization of total well-being, and derived what is known as the Sidgwick-Meade rule according to which, at the optimum, the marginal utility of consumption is equal to average well-being level per unit of consumption.

The Sidgwick-Meade rule has been obtained in what Dasgupta (2005) named the *genesis problem*, where there is a total amount of a consumption good available. The problem is to fix the optimal number of individuals such that, sharing the consumption good equally among them, brings the largest social welfare. Asheim and Zuber (2014) studied this similar problem using a more general rank-dependent generalized utilitarian criterion and provided a more general condition than the Sidgwick-Meade rule for optimal population.

Of course, as argued by Dasgupta (2005), the genesis problem is not be the most interesting problem for optimal population theory as it neglects that the resource, for instance a natural resource, has to be used by several successive generations. It may thus regenerate but only if something is left to the next generation. Nerlov, Razin and Sadka (1985) have studied a simple two-periods extension of the genesis model, with a non-renewable resource, assuming that parents have altruistic sentiments towards their children. Spiegel (1993) extended their analysis to the standard maximin case (without the adjustment for population size that we proposed in Section 3).[6] In this paper I study extensions of their framework that allow for more general production processes using a natural resource that may be renewable or non-renewable.

The framework hence involves a current generation, whose size N is exogenously given. This generation has a consumption c and derives utility $u(c)$. There is also a future generation, whose size is rN. The number r is the reproduction rate to be chosen. This future generation's consumption level is d (and utility level $u(d)$). The problem thus

[5] A recent questionnaire-experimental study shows that many people actually do not share the view that only average well-being matters for the optimum population, but that the size of the population itself is important (Spears, 2017).

[6] There is also a very large literature focusing on infinite-horizon models of economic growth, for instance Dasgupta (1969), Palivos and Yip (1993), Razin and Yuen (1995), Boucekkine, Fabbri and Gozzi (2014) and Lawson and Spears (2018). This literature studies how per capita well-being and population size is optimized, focusing on total and average utilitarianism. However, they focus on within-generation optimal population, while I focus on the optimal total population size across generations. The results below cannot easily be translated in their context.

consists in choosing the values of c, d and r. To simplify, r is treated as a continuous variable.

An allocation can therefore be seen as a triplet $x = (r, c, d)$. Using the notation of Section 2, let denote $n(x) = (1 + r)N$. Let denote $F \subset \mathbb{R}^3_+$ the set of feasible allocation. At this stage, I do not impose any restriction of F that may result from the use of natural and non-natural resources, and may involve costs from child rearing.

To determine the optimal allocation, assume that there exists an SWO $\succsim$ on $\mathbb{R}^3_+$. I follow Blackorby, Bossert and Donaldson (2001) and assume that the SWO has the following form:

$$x \succsim y \iff V\Big(n(x), \Xi(x)\Big) \geq V\Big(n(y), \Xi(y)\Big),$$

with function Ξ a within-generation equally-distributed equivalent function that represent the ordering of allocations for a given total population size, and function V an aggregator function that combines the equally-distributed equivalent welfare (henceforth EDEW) Ξ and population size. Function V describes how population size and equivalent per capita welfare are traded-off and thus embodies the population ethics views of the decision maker.

For instance, the utilitarian EDEW (expressed in terms of utility) is

$$\Xi^U(r, c, d) = \tfrac{1}{1+r}u(c) + \tfrac{r}{1+r}u(d) \tag{1}$$

while the maximin EDEW is

$$\Xi^E(r, c, d) = \min\big\{u(c), u(d)\big\}. \tag{2}$$

In the utilitarian case, two prominent aggregator functions have been considered. Function $V^T(n, e) = n \times e$ delivers the total utilitarian (or Benthamite) social welfare function:

$$W^{TU}(x) = V^T\Big(n(x), \Xi^U(x)\Big) = Nu(c) + rNu(d).$$

Function $V^A(n, e) = e$ delivers the average utilitarian (or Millian) social welfare function:

$$W^{AU}(x) = V^A\Big(n(x), \Xi^U(x)\Big) = \tfrac{1}{1+r}u(c) + \tfrac{r}{1+r}u(d).$$

Given the form of the SWOs, there are two questions that may be asked. First, given a specific view about how resources should be allocated in a population of a given size (as embodied in function Ξ), how does the population ethics view, that is the shape of function V, influence the optimal population growth rate? This question was raised in the specific case of the total utilitarian and average utilitarian views described above by Edgeworth (1925). Edgeworth conjectured that the socially optimal rate of population growth must be larger for a total than for an average utilitarian social welfare function, and Nerlov, Razin and Sadka (1985) actually proved it in a specific model. In Section 4, I show that this is a special case of a more general result, which is not restricted to utilitarianism nor to a specific economic model of resource allocation, and that we can compare more aggregator functions.

A different question, that has attracted less attention, is to understand whether (and how) the ethical view about the allocation for a given population size influences optimal population growth. Does utilitarianism require a larger population than egalitarianism,

for a given population ethical view as represented by the aggregator function V? It turns out that this so for the average view and for low levels of resources (Section 4), but not in general.

Population ethical views and optimal population size: a general result

In this section, I study the implications of population ethics views as represented by the aggregator function V on optimal population growth r. I thus take the ethical view about redistribution within a population (embodied in the EDEW Ξ) as given (it may a utilitarian or egalitarian or any other kind of view).

To study the question, let us consider a family $(V_\theta)_{\theta\in\Theta}$ of aggregator functions, with Θ a set of parameters, which is a subset of $\mathbb{R}$ (for instance $\Theta = [0,1]$). Taking Ξ as given, this delivers a family of SWOs $\succsim_\theta^\Xi$:

$$x \succsim_\theta^\Xi y \iff V_\theta\Big(n(x), \Xi(x)\Big) \geq V_\theta\Big(n(y), \Xi(y)\Big).$$

with $\theta \in \Theta$ a specific member of this family.

To compare the implications of members of family $(V_\theta)_{\theta\in\Theta}$ for optimal population size, I make the following assumption.

Assumption 1 *There exists a threshold $\gamma \in \mathbb{R}_+$ such that family $(V_\theta)_{\theta\in\Theta}$ satisfies the following conditions for any $\theta \in \Theta$:*

Regularity:

- *for any $n > n'$ and $e > \gamma$, $V_\theta(n,e) > V_\theta(n',e)$;*
- *for any n, n' and $e \leq \gamma < e'$, $V_\theta(n',e') > V_\theta(n,e)$;*
- *for any n and $e > e'$, $V_\theta(n,e) > V_\theta(n,e')$.*

Sorting: *for any $n > n'$ and $\gamma < e < e'$, if $V_\theta(n,e) \geq V_\theta(n',e')$ then $V_{\theta'}(n,e) > V_{\theta'}(n',e')$ for all $\theta' > \theta$.*

The regularity condition means that the social welfare function is increasing in both EDEW and population size (provided welfare is high enough). It also implies that population cannot compensate for welfare if welfare is below a certain threshold: if $\gamma = 0$, this corresponds to what Blackorby, Bossert and Donaldson (2005) name the 'Priority to lives worth living' (we prefer populations with equal positive well-being to populations with equal negative well-being).

The sorting condition means that the family in well-ordered in terms of how the SWOs trade-off population and EDEW (at least for large enough well-being levels): a higher θ means that we want to give up more on well-being to increase population.

Here are two examples of families using the population-weighted approach and satisfying Assumption 1:

- For all $\theta \in [0,1]$, all $n \in \mathbb{N}$ and all $e \in \mathbb{R}$,

$$V_\theta^p(n,e) = \begin{cases} e & \text{if } e \leq 0; \\ n^\theta \cdot e & \text{if } e > 0. \end{cases} \tag{3}$$

- For all $\theta \in [0,1)$, all $n \in \mathbb{N}$ and all $e \in \mathbb{R}$, $V_\theta^g(n,e) = \begin{cases} e & \text{if } e \leq 0; \\ \frac{1-\theta^n}{1-\theta} \cdot e & \text{if } e > 0. \end{cases}$

The following general result is true for all EDEW Ξ.

Proposition 3 *Consider a family $(V_\theta)_{\theta\in\Theta}$ that satisfies Assumption 1 for some threshold $\gamma \in \mathbb{R}_{++}$ and a feasible set F such that, for some $\bar{x} \in F$, $\Xi(\bar{x}) > \gamma$. For each $\theta \in \Theta$ assume that there exists an allocation $x_\theta^\star \in F$ such that $V_\theta\Big(n(x_\theta^\star), \Xi(x_\theta^\star)\Big) \geq V_\theta\Big(n(y), \Xi(y)\Big)$ for all $y \in F$.*
Then, for any $\theta' > \theta$, $n(x_{\theta'}^\star) \geq n(x_\theta^\star)$.

Proof. By definition of $x_\theta^\star$ and $x_{\theta'}^\star$, we have $V_\theta\Big(n(x_\theta^\star), \Xi(x_\theta^\star)\Big) \geq V_\theta\Big(n(\bar{x}), \Xi(\bar{x})\Big)$ and $V_{\theta'}\Big(n(x_{\theta'}^\star), \Xi(x_{\theta'}^\star)\Big) \geq V_{\theta'}\Big(n(\bar{x}), \Xi(\bar{x})\Big)$. Hence, by the regularity condition in Assumption 1, $\Xi(x_\theta^\star) > \gamma$ and $\Xi(x_{\theta'}^\star) > \gamma$.

Assume that $n(x_{\theta'}^\star) < n(x_\theta^\star)$.

If $\Xi(x_{\theta'}^\star) \leq \Xi(x_\theta^\star)$, given that $\Xi(x_\theta^\star) > \gamma$, by the the regularity condition in Assumption 1, we would have $V_{\theta'}\Big(n(x_\theta^\star), \Xi(x_\theta^\star)\Big) > V_{\theta'}\Big(n(x_{\theta'}^\star), \Xi(x_{\theta'}^\star)\Big)$. This is a contradiction of the definition of $x_{\theta'}^\star$.

If $\Xi(x_{\theta'}^\star) > \Xi(x_\theta^\star)$, we have $n' = n(x_{\theta'}^\star) < n(x_\theta^\star) = n$, $e' = \Xi(x_{\theta'}^\star) > \Xi(x_\theta^\star) = e > \gamma$ and $V_\theta(n,e) \geq V_\theta(n',e')$ by definition of $x_\theta^\star$. By the sorting condition in Assumption 1, this implies $V_{\theta'}\Big(n(x_\theta^\star), \Xi(x_\theta^\star)\Big) > V_{\theta'}\Big(n(x_{\theta'}^\star), \Xi(x_{\theta'}^\star)\Big)$. This is again a contradiction of the definition of $x_{\theta'}^\star$. ■

Proposition 3 confirms Edgeworth's conjecture. Indeed, if we take $\Xi = \Xi^U$ and consider the V_θ^p (Eq. 3), with $\theta \in [0,1]$ and $\gamma = 0$, we clearly see that the total utilitarian view corresponds to $\theta = 1$ while the average utilitarian view corresponds to $\theta = 0$. Hence the former induces a larger population than the latter, whatever the economic model inducing the feasible set F.

But Proposition 3 also generalizes Edgeworth's conjecture in several ways. First, and as mentioned just above, the result is true for any feasible set F. Second, the result does not only make it possible to compare a "total" and an "average" view, but many intermediary views corresponding to $0 < \theta < 1$. Last, the result is not restricted to utilitarianism and may apply to other ethical views, including egalitarianism. In Theorem 2, we have shown that population-adjusted egalitarian criteria have appealing properties. In the framework of this section, they are represented by

$$W_\theta^E(x) = \Pi_{\kappa^\theta}\Big((1+r)N, \min\{u(c), u(d)\}\Big),$$

with κ^θ a non-decreasing sequence with $\kappa_1^\theta = 1$. If $u(z) = 0$ for some $z \in \mathbb{R}_{++}$ and the sorting condition is satisfied, we obtain families of social welfare functions that satisfy Assumption 1. We can then apply Proposition 3.

Optimal population size: egalitarianism vs. utilitarianism

A key question has not been much addressed in the literature: does the egalitarian view imply larger or smaller populations than the utilitarian view? In the genesis problem that

has often been studied (Dasgupta, 2005), there is no distinction between utilitarianism and egalitarianism because both of them recommend that consumption should be equalized among individuals, so that they yield the same equally-distributed equivalent for a given population size. To be able to distinguish the two views, we should consider cases where an unequal distribution of consumption (and hence utility) between generations may be optimal. This can be obtained in a simple model involving a renewable natural resource and/or technological progress.

This very simple model specifies the form of the feasible set F and of the utility function u. I assume that there is an initial stock of natural resource Ω. We can use an amount ω_1 of resource in the first period to produce the consumption good, so that $Nc \leq A\,\omega_1$, where $A \in \mathbb{R}_{++}$ is the total factor productivity. The remaining stock of the natural resource will reproduce so that there is an amount $(1+\phi)(\Omega-\omega_1)$ available to the next generation, where $\phi \in \mathbb{R}_+$ is the regeneration rate ($\phi = 0$ corresponds to non-renewable resource). The second generation (of size rN) can use $\omega_2 \leq (1+\phi)(\Omega-\omega_1)$ to produce the consumption good, so that $rNd \leq (1+g)\,A\,\omega_2$, where $g \in \mathbb{R}_+$ is the rate of technological progress (assumed exogenous).

Rewriting the constraints, we have $\omega_1 + \frac{\omega_2}{1+\phi} \leq \Omega$, $c \leq A\,\frac{\omega_1}{N}$ and $r(1+g)^{-1}\,d \leq A\,\frac{\omega_2}{N}$. This gives the following aggregate constraint:

$$c + \tfrac{r}{1+\tilde{g}}d \leq \tilde{\omega},$$

with $\tilde{g} = (1+\phi)(1+g) - 1$ a 'total' productivity growth rate (that combines resource growth and pure technology growth) and $\tilde{\omega} = A\,\frac{\Omega}{N}$ an index of per capita initial resource (that corresponds to the maximum per capita consumption in the first generation if the natural resource is completely exhausted in the first period). Hence the feasible set is $F = \left\{(r,c,d) | c + \frac{r}{1+\tilde{g}}d \leq \tilde{\omega}\right\}$.

The aim of this section is to compare the egalitarian and utilitarian approaches for a given aggregator function. To fit with the population-adjusted egalitarian approach introduced in this paper, let us consider the aggregator function V_θ^g that has already been mentioned above (Eq. 3).

I assume that individual well-being is measured by an iso-elastic function of consumption and I normalize consumption so that a consumption level of 1 corresponds to the neutral level of well-being, i.e. $u(1) = 0$:[7]

$$u(c) = \tfrac{c^{1-\varepsilon}}{1-\varepsilon} - \tfrac{1}{1-\varepsilon}, \qquad \varepsilon > 0.$$

Thus, the utilitarian EDEW is:

$$\Xi^U(r,c,d) = \tfrac{1}{1+r}\tfrac{c^{1-\varepsilon}}{1-\varepsilon} + \tfrac{r}{1+r}\tfrac{d^{1-\varepsilon}}{1-\varepsilon} - \tfrac{1}{1-\varepsilon}.$$

The egalitarian EDEW is $\Xi^E(r,c,d) = \frac{(\min\{c,d\})^{1-\varepsilon}-1}{1-\varepsilon}$.

[7] The normalization of the neutral level of consumption to 1 is without loss of generality, given that the general definition with a utility of 0 at consumption level γ is $u(c) = \frac{c^{1-\varepsilon}}{1-\varepsilon} - \frac{\gamma^{1-\varepsilon}}{1-\varepsilon} = \gamma^{1-\varepsilon}\left(\frac{(c/\gamma)^{1-\varepsilon}}{1-\varepsilon} - \frac{1}{1-\varepsilon}\right)$. The multiplication by the positive constant $\gamma^{1-\varepsilon}$ does not change the social ranking.

We look for the optimal allocation according to the utilitarian and egalitarian SWOs, that is a solution to the utilitarian and egalitarian problems. The utilitarian problem is:

$$\max_{(r,c,d)\in F} V_\theta^g \left((1+r)N, \tfrac{1}{1+r}\tfrac{c^{1-\varepsilon}}{1-\varepsilon} + \tfrac{r}{1+r}\tfrac{d^{1-\varepsilon}}{1-\varepsilon} - \tfrac{1}{1-\varepsilon}\right). \tag{4}$$

The egalitarian problem is:

$$\max_{(r,c,d)\in F} V_\theta^g \left((1+r)N, \tfrac{(\min\{c,d\})^{1-\varepsilon}}{1-\varepsilon} - \tfrac{1}{1-\varepsilon}\right). \tag{5}$$

Denote $(r_\theta^{E\star}, c_\theta^{E\star}, d_\theta^{E\star}) \in F$ (resp. $(r_\theta^{U\star}, c_\theta^{U\star}, d_\theta^{U\star}) \in F$) a solution to the egalitarian problem (5) (resp. a solution to the utilitarian problem (4)). To solve these problems, we can proceed in two steps. First, for each possible population growth level r, we find the optimal consumption level for the first and second generations and compute the equally-distributed equivalent. Then we optimize with respect to r.

Optimum EDEW and the average view

Routine reasoning implies that, for a given r, the optimal egalitarian allocation requires $c = d$. The resource constraint $c + \frac{r}{1+\tilde{g}}d \leq \tilde{\omega}$ (given that utility is increasing in consumption) implies that

$$\tfrac{1+\tilde{g}+r}{1+\tilde{g}}c = c + \tfrac{r}{1+\tilde{g}}d = \tilde{\omega},$$

so that $c = \frac{1+\tilde{g}}{1+\tilde{g}+r}\tilde{\omega}$. The optimum egalitarian EDEW for a population growth r is therefore:

$$EDEW^E(r) = \left(\tfrac{1+\tilde{g}}{1+\tilde{g}+r}\right)^{1-\varepsilon} \tfrac{\tilde{\omega}^{1-\varepsilon}}{1-\varepsilon} - \tfrac{1}{1-\varepsilon}. \tag{6}$$

This is clearly decreasing in r: increasing population size necessarily comes to a cost, which is a decrease in the egalitarian EDEW.

Similarly, is is easy to show that, for a given r, the optimal utilitarian allocation is such that $d = (1+\tilde{g})^{\frac{1}{\varepsilon}}c$. The resource constraint implies that

$$\left(1 + r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\right)c = c + \tfrac{r}{1+\tilde{g}}d = \tilde{\omega},$$

so that $c = \left(1 + r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\right)^{-1}\tilde{\omega}$ and $d = (1+\tilde{g})^{\frac{1}{\varepsilon}}\left(1 + r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\right)^{-1}\tilde{\omega}$. The optimum utilitarian EDEW for a population growth r is therefore:

$$\begin{aligned} EDEW^U(r) &= \left[\tfrac{1+r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}}{1+r}\right]\left(1 + r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\right)^{-(1-\varepsilon)} \tfrac{\tilde{\omega}^{1-\varepsilon}}{1-\varepsilon} - \tfrac{1}{1-\varepsilon} \\ &= \left(1 + r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\right)^{\varepsilon}(1+r)^{-1}\tfrac{\tilde{\omega}^{1-\varepsilon}}{1-\varepsilon} - \tfrac{1}{1-\varepsilon} \end{aligned} \tag{7}$$

Contrary to the egalitarian case, this utilitarian EDEW may not be always decreasing. Indeed, if the total productivity growth rate is large enough, the utilitarian EDEW may

be first increasing and then decreasing so that there exists a unique non zero population growth rate r that maximizes $EDEW^U$.[8]

The behavior of the EDEW at the optimal allocation for a given population growth gives a first result, in the case where $\theta = 0$ in Eq.(3), that is when no additional weight is given to larger population. We call this case the average view, given that it corresponds to average utilitarianism when a utilitarian criterion is used. Recall that the optimal levels of population growth when $\theta = 0$ are denoted $r_0^{E\star} = 0$ and $r_0^{U\star}$.

Proposition 4 *For the average view,* $r_0^{E\star} = 0$ *while* $r_0^{U\star} > 0$ *if* $\ln(1+\tilde{g}) > \frac{\varepsilon}{\varepsilon-1}\ln(\varepsilon)$ *(and* $r_0^{U\star} = 0$ *otherwise).*

Prop. 4 states that egalitarianism always recommends no population growth in the average view, in order to have the largest possible level of minimal welfare. This is not true of utilitarianism that may recommend population growth if it contributes to increasing average welfare. When the total productivity growth rate is high, it is possible to have a rich enough second generation without hurting the first one too much, so as to increase average welfare.

The general case

Let us now consider the case where $\theta > 0$ in Eq. (3): we multiply EDEW by an increasing function of population size (provided that EDEW is positive). Using the results from Section 4, we can rewrite the egalitarian objective function in problem (5) as follows:

$$U_\theta^E(r) = V_\theta^p\left((1+r)N, \left(\tfrac{1+\tilde{g}}{1+\tilde{g}+r}\right)^{1-\varepsilon} \tfrac{\tilde{\omega}^{1-\varepsilon}}{1-\varepsilon} - \tfrac{1}{1-\varepsilon}\right).$$

If $\tilde{\omega} \leq 1$, then $\left(\frac{1+\tilde{g}}{1+\tilde{g}+r}\right)^{1-\varepsilon} \frac{\tilde{\omega}^{1-\varepsilon}}{1-\varepsilon} < \frac{1}{1-\varepsilon}$ (EDEW is always negative) and $U_\theta^E(r) = EDEW^E(r)$ for all $r > 0$. We know from Section Section 4 that maximizing $EDEW^E$ always yield $r_\theta^{E\star} = 0$. In that case, the utilitarian approach will always induces an (at least weakly) larger population. I will thus focus on cases where $\tilde{\omega} > 1$.

Proposition 5 *If* $\varepsilon > 1$, $\tilde{\omega} > 1$ *and* $\ln(1+\tilde{g}) \leq 1$ *then there exists unique solutions* $r_\theta^{E\star}$ *and* $r_\theta^{U\star}$.

Furthermore, there exist $1 < \underline{\omega} < \bar{\omega}$ *such that:*

1. *If* $\tilde{\omega} \leq \underline{\omega}$, *then* $r_\theta^{U\star} = r_\theta^{E\star} = 0$,

[8] The derivative of $EDEW^U$ with respect to r is indeed:

$$(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\varepsilon\left(1+r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\right)^{\varepsilon-1}(1+r)^{-1}\tfrac{\tilde{\omega}^{1-\varepsilon}}{1-\varepsilon} - \left(1+r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\right)^{\varepsilon}(1+r)^{-2}\tfrac{\tilde{\omega}^{1-\varepsilon}}{1-\varepsilon}$$

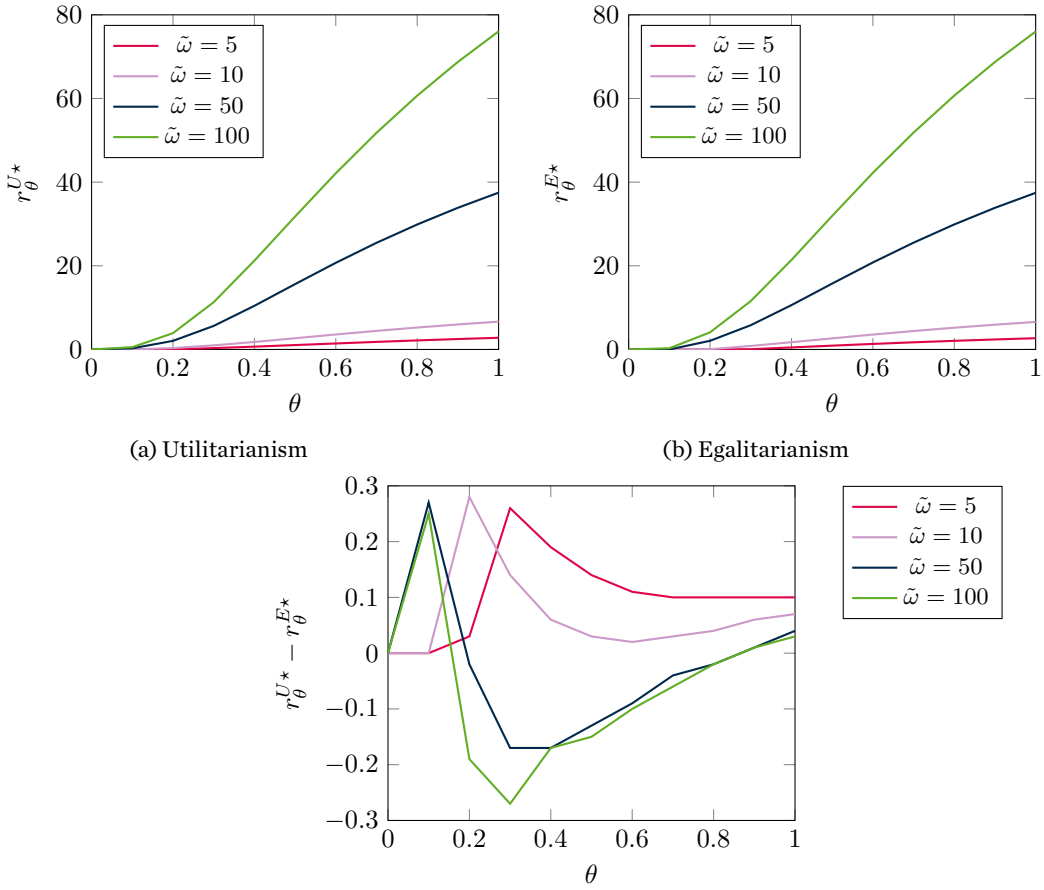
$$= \tfrac{(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\varepsilon(1+r)-1+r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}}{1-\varepsilon}\left(1+r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\right)^{\varepsilon-1}(1+r)^{-2}\,\tilde{\omega}^{1-\varepsilon},$$

which depends on the sign of $\frac{(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\varepsilon-1}{1-\varepsilon} - (1+\tilde{g})^{\frac{1}{\varepsilon}-1}r$. This sign is eventually negative but may first be positive whenever $\frac{(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\varepsilon-1}{1-\varepsilon} > 0$, or equivalently $\ln(1+\tilde{g}) > \frac{\varepsilon}{\varepsilon-1}\ln(\varepsilon)$.

2. *if* $\underline{\omega} < \tilde{\omega} \leq \bar{\omega}$, *then* $r_\theta^{U\star} > r_\theta^{E\star} = 0$.

Prop. 5 shows that utilitarianism may recommend a larger population growth than egalitarianism when resources are low. Actually, it recommends a strictly larger population growth whenever the initial level of resources is at an intermediate level (not lower than $\underline{\omega} > 1$ but not too large either).

This seemingly hints at a potential general result that utilitarianism always recommend larger population growth. However, there is no such general result as illustrated in Figure 1.



(a) Utilitarianism

(b) Egalitarianism

(c) Difference

Figure 1: Optimal population for different levels of initial resources

Figure 1 provides numerical results of optimal population growth for various levels of initial resources $\tilde{\omega}$ in the case where $\varepsilon = 1.1$ and $\tilde{g} = 1$ (so that $\ln(1+\tilde{g}) < 1$). Panel (a) of Figure 1 gives optimal population growth $r_\theta^{U\star}$ for the utilitarian case and Panel (b) gives optimal population growth $r_\theta^{E\star}$ for the egalitarian case. The main first insight is that $r_\theta^{U\star}$

and $r_\theta^{E\star}$ exhibit very similar patterns when θ and $\tilde{\omega}$ vary. In particular, Figure 1 illustrates the result in Prop. 3: an increase in θ always yield a larger optimal population both for the utilitarian and egalitarian approaches. Similarly, an increase in initial resources $\tilde{\omega}$ always increases optimal population sizes.[9] It is perhaps more surprising that the exact levels of optimal population growth seems very close in the two approaches.

To look closer into the exact ranking of utilitarianism and egalitarianism in terms of optimal population growth, Panel (c) of Figure 1 draws the difference $(r_\theta^{U\star} - r_\theta^{E\star})$. It shows that the difference is usually very small, even when population growth is large. More strikingly compared to Prop. 5, Panel (c) shows that there are cases where the difference is negative, that is the egalitarian optimal population growth is larger than the utilitarian one. This happens when $\tilde{\omega}$ is large enough ($\tilde{\omega} = 50$ and $\tilde{\omega} = 100$) and θ has intermediates values (in the range $[0.2, 0.8]$). In any case, this proves that there is no hope to obtain general results regarding the relative utilitarian and egalitarian optimal population sizes.

5 Concluding remarks

In this paper, I have provided arguments why egalitarianism may be an attractive view in intergenerational models where resources have to be allocated between successive generations, whose number and size may be affected by the policy. If we want to limit the sacrifices made by the current generation for the sake of the future, or if we want to promote small sacrifices that benefit many future people, egalitarianism is appealing. I have also provided a new class of egalitarian criteria that may be used to compare populations with different sizes, while embodying many views about population ethics.

Applying these criteria to the question of optimal future population size, when we use renewable or non-renewable resources, I have exhibited a condition (the sorting condition) that orders population ethics views in terms of the optimal population size they recommend. We can thus define a 'total egalitarian' and an 'average egalitarian' view (much like the 'total utilitarian' and 'average utilitarian' views) and show that the former induces more population than the latter. But we can also define many intermediate views that are worth studying in more details. In a simple model I have also showed that egalitarianism may recommend less population growth than utilitarianism in specific cases (average view and low level of resources) but that this is not true in general. It would be interesting to study the implications of Population-adjusted egalitarianism in more complex models. This will be the topic of future work.

References

Arrhenius G. (2000). "An impossibility theorem for welfarist axiologies", *Economics & Philosophy* **16**, 247–266.

Arrhenius G. (2019). *Population Ethics – The Challenge of Future Generations*. Forthcoming.

[9] A proof that this is true in general in the present setting when $\varepsilon > 1$ and $\ln(1 + \tilde{g}) < 1$ is available upon request.

Asheim G.B., Mitra T. ,Tungodden B. (2007). "A new equity condition for infinite utility streams and the possibility of being Paretian." In J. Roemer and K. Suzumura (Eds.) *Intergenerational Equity and Sustainability*, pp. 55–68. Basingstoke: Palgrave Macmillan.

Asheim G.B., Zuber S. (2014). "Escaping the repugnant conclusion: Rank-discounted utilitarianism with variable population," *Theoretical Economics* **9** 629–650.

Barberá S., Jackson M. (1988). "Maximin, leximin, and the protective criterion: characterizations and comparisons," *Journal of Econonomic Theory* **46**, 34–44.)

Blackorby C., Bossert W., Donaldson D. (1996). "Leximin population ethics," *Mathematical Social Sciences* **31**, 115–131.

Blackorby C., Bossert W., Donaldson D. (2001). "Population ethics and the existence of value functions," *Journal of Public Economics* **82**, 301–308.

Blackorby C., Bossert W., Donaldson D. (2005). *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*. Cambridge: Cambridge University Press.

Blackorby C., Donaldson D. (1982). "Ratio-scale and translation-scale full interpersonal comparability without domain restrictions: Admissible social-evaluation functions," *International Economic Review* **23**, 249–268.

Bossert W. (1990). "Maximin welfare orderings with variable population size," *Social Choice and Welfare* **7**, 39–45.

Boucekkine R., Fabbri G., Gozzi F. (2014). "Egalitarianism under population change: Age structure does matter," *Journal of Mathematical Economics* **55**, 86–100.

Carlson E. (1998). "Mere addition and two trilemmas of population ethics," *Economics and Philosophy* **14**, 283–306.

Dasgupta P.S. (1988). "On the concept of optimum population," *Review of Economic Studies* **36**, 295–318.

Dasgupta P.S. (2005). "Regarding optimum population", *Journal of Political Philosophy* **13**, 414–442.

d'Aspremont C., Gevers L. (2002). "Social welfare functionals and interpersonal comparability." In K.J. Arrow, A.K. Sen and K. Suzumura (Eds.) *Handbook of Social Choice and Welfare–vol. 1*, pp. 459–541. Amsterdam: Elsevier.

Edgeworth F. Y. (1925). *Papers Relating to Political Economy - Vol. 3*. London: MacMillan.

Fleurbaey M., Maniquet F. (2011). *A Theory of Fairness and Social Welfare*. Cambridge: Cambridge University Press.

Fleurbaey M., Tungodden B. (2010). "The tyranny of non-aggregation versus the tyranny of aggregation in social choices: a real dilemma," *Economic Theory* **44**, 399–414.

Hammond P.J. (1976). "Equity, Arrow's conditions, and Rawls' difference principle," *Econometrica* **44**, 793–804.

Hartwick J.M. (1977). "Intergenerational equity and the investing of rents from exhaustible resources," *American Economic Review* **67**, 972–974.

Intergovernmental Panel on Climate Change (2015a). *Climate Change 2014: Impacts, Adaptation, and Vulnerability*. Cambridge: Cambridge University Press.

Intergovernmental Panel on Climate Change, (2015b). *Climate Change 2014: Mitigation of Climate Change*. Cambridge: Cambridge University Press.

Lauwers L. (1997). "Rawlsian equity and generalized utilitarianism with an infinite population," *Econonomic Theory* **9**, 143–150.

Lawson N., Spears D. (2018). "Optimal population and exhaustible resource constraint," *Journal of Population Economics* **31**, 295–335.

Malthus, T.R. (1798). *An Essay On The Principle Of Population*. London: Penguin Books.

Meade J.E. (1955). *Trade and Welfare*. Oxford: Oxford University Press.

Mitra T. (2002). "Intertemporal equity and efficient allocation of resources," *Journal of Econonomic Theory* **107**, 356–376.

Nerlov M., Razin A., Sadka E. (1985). "Population Size: Individual Choice and Social Optima," *The Quarterly Journal of Economics* **100**, 321–334.

Palivos T., Yip C.K. (1993). "Optimal population size and endogenous growth," *Economics Letters* **41**, 107–110.

Parfit D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.

Razin A., Yuen C.W. (1995). "Utilitarian tradeoff between population growth and income growth," *Journal of Population Economics* **8**, 81–87.

Rawls J. (1971). *A Theory of Justice*. Oxford: Oxford University Press.

Roberts K.W.S. (1980). "Interpersonal comparability and social choice theory," *Review of Economic Studies* **47**, 421–446.

Sen A.K. (1980). "Equality of what." In S. McMurrin (Ed.) *Tanner Lectures on Human Values–vol. I*. Cambridge: Cambridge University Press.

Sen A.K. (1986). "Social choice theory." In K.J. Arrow, M.D. Intriligator (Eds.) *Handbook of Mathematical Economics–vol. III*, pp. 1073–1181. Amsterdam: Elsevier.

Solow R.M. (1974). "Intergenerational equity and exhaustible resources," *Review of Economic Studies* **41**, 29–45.

Spears D. (2017). "Making people happy or making happy people? Questionnaire-experimental studies of population ethics and policy," *Social Choice and Welfare* **49**, 145–169.

Spiegel Y. (1993). "Rawlsian optimal population size," *Journal of Population Economics* **6**, 363–373.

Wicksell K. (1893). *Value, Capital and Rent*. London: George Allen& Unwin Ltd. [1954 edition]

Withagen C., Asheim G.B. (1998). "Characterizing sustainability: The converse of Hartwick's rule," *Journal of Economic Dynamics and Control* **23**,159–165.

A Appendix A: Proofs

Preliminary result

For any $x \in X$ and any $k \in \mathbb{N}$, let $k \star x$ denote $y \in X$ such that $n(y) = kn(x)$ and, for any $\ell \in \{1, \cdots, k\}$ and any $i \in \{1, \cdots, n(x)\}$, $y_{(\ell-1)+i} = x_i$. This is a k-replica of x.

Consider the following principle, which is very common in the literature. It states that replicating several times the same welfare distributions should not alter social judgements.

Replication Invariance. For all $n \in \mathbb{N}$, all $x, y \in \mathbb{R}^n$ and all $k \in \mathbb{N}$, $k \star x \succsim k \star y$ if and only if $x \succsim y$.

This principle is implied by Consistency.

Lemma 1 *If an SWO $\succsim$ on X satisfies Consistency then it satisfies Replication Invariance.*

Proof. Assume that the SWO $\succsim$ on X satisfies Consistency. Consider any $n \in \mathbb{N}, x, y \in \mathbb{R}^n$ and $k \in \mathbb{N}$. Assume that $x \succsim y$. Then by k applications of Consistency $k \star x \succsim k \star y$.

Conversely assume that $k \star x \succsim k \star y$ but $x \prec y$. By k applications of Consistency, we should have $k \star x \prec k \star y$, which is a contradiction. ■

Proof of Proposition 1

Assume that the SWO $\succsim$ satisfies Suppes-Sen, Pigou-Dalton and Consistency (and thus Replication invariance by Lemma 1). If $n = 1$, for any $x, y \in \mathbb{R}$ if $x_{[1]} > y_{[1]}$ then $x \succ y$ by Suppes-Sen. Below we focus on cases where $n > 1$.

Proof of statement 1.

Assume that $\succsim$ also satisfies Limited sacrifice for the rich future.

Consider any $n \in \mathbb{N} \setminus \{1\}$ and $x, y \in \mathbb{R}^n$, such that $x_{[1]} > y_{[1]}$. If $x_{[1]} > y_{[n]}$, then $x \succ y$ by Suppes-Sen. If $x_{[1]} \leq y_{[n]}$, consider the real numbers $a, b, c \in \mathbb{R}$ such that $y_{[1]} < a < b < x_{[1]} \leq y_{[n]} < c$.

Define $\alpha = b - a$ and let β be the corresponding term in the statement of Limited sacrifice for the rich future. Let $0 < \gamma < \beta$ and $m \in \mathbb{N}$ be such that $b = c - m\gamma$. Consider a collection of allocations $(w^0, w^1, \cdots, w^m)$ such that $w^0 = \big((a)_m, (c)_{m(n-1)}\big)$, $w^m = (b)_{mn}$ and for each $k \in I_{m-1}$ $w^k = \big((b)_k, (a)_{m-k}, (c - k\gamma)_{m(n-1)}\big)$.

By Limited sacrifice for the rich future, $\big((b)_1, (c - k\gamma)_{m(n-1)}\big) \succsim \big((a)_1, (c - (k - 1) \cdot \gamma)_{m(n-1)}\big)$ for all $k \in I_m$. Denoting $z^1 = (a)_{m-1}$ and $z^k = \big((b)_{k-1}, (a)_{m-k}\big)$, by Consistency this implies

$$w^k = \big(z^k, (b)_1, (c - k\gamma)_{m(n-1)}\big) \succsim \big(z^k, (a)_1, (c - (k-1) \cdot \gamma)_{m(n-1)}\big) = w^{k-1}$$

for all $k \in I_m$. By transitivity, we obtain $w^m \succsim w^0$ and by Replication invariance $b_n \succsim \big((a)_1, (c)_{n-1}\big)$.

But, given that $b < x_{[1]}$, Suppes-Sen implies that $x \succ (b)_n$. Similarly, given that $y_{[1]} < a$ and $y_{[n]} < c$, Suppes-Sen implies that $\big((a)_1, (c)_{(n-1)}\big) \succ y$. By transitivity, $x \succ y$.

Proof of statement 2.

Assume that $\succsim$ also satisfies Limited sacrifice for the long future.

Step 1: there exists $k \in \mathbb{N}$ such that for any $a, b, c, d \in \mathbb{R}$ with $a < b \leq c < d$ and any $n \geq k$, $\big((c)_1, (b)_n\big) \succ \big((d)_1, (a)_n\big)$.

Let k be the number in the statement of Limited sacrifice for the long future. Consider any $a, b, c, d \in \mathbb{R}$ such that $a < b \leq c < d$ and any $n \geq k$. Let $m \in \mathbb{N}$ and $\theta \in \mathbb{R}_{++}$ be such that $d - m\theta = a$ and $\theta \leq \gamma$, where γ is the number in the statement of Limited sacrifice for the long future. Denote $\varepsilon = \frac{b-a}{m}$.

By Limited sacrifice for the long future, $\big((a + (\ell+1)\cdot\varepsilon)_n, (d - (\ell+1)\cdot\theta)_1\big) \succsim \big((a + \ell\cdot\varepsilon)_n, (d - \ell\cdot\theta)_1\big)$ for all $\ell = 0, \cdots, m-1$, so that, by transitivity, $\big((b)_n, (c)_1\big) \succsim \big((a)_n, (d)_1\big)$.[10]

Step 2: there exists $k \in \mathbb{N}$ such that for any $a, b, c \in \mathbb{R}$ with $a < b \leq c$ and any $n \geq k$ and $m \in \mathbb{N}$, $\big((c)_1, (b)_{m+n-1}\big) \succsim \big((c)_m, (a)_n\big)$.

For any $a, b, c \in \mathbb{R}$ with $a < b \leq c$ and $m \in \mathbb{N}$, consider a collection of real numbers $(d^1, \cdots, d^m)$ such that $d^1 = b$, $d^m = a$ and $d^1 > d^2 > \cdots > d^m$. Let $n \geq k$, with $k \in \mathbb{N}$ the number in Step 1. Consider a collection of allocations $(w^1, \cdots, w^m)$ such that, for each $k \in I_m$ $w^k = \big((c)_k, (d^k)_{m+n-k}\big)$.

By Step 1, for each $k \in I_{m-1}$, we have $(d^k)_{m+n-k} \succsim \big((c)_1, (d^{k+1})_{m+n-k-1}\big)$. By Consistency, this implies $\big((c)_k, (d^k)_{m+n-k}\big) \succsim \big((c)_k, (c)_1, (d^{k+1})_{m+n-k-1}\big)$, which can be written $w^k \succsim w^{k+1}$. By transitivity, this implies that $\big((c)_1, (b)_{m+n-1}\big) = w^1 \succsim w^m = \big((c)_m, (a)_n\big)$.

Step 3: Conclusion.

Consider any $n \in \mathbb{N} \setminus \{1\}$ and $x, y \in \mathbb{R}^n$, such that $x_{[1]} > y_{[1]}$. If $x_{[1]} > y_{[n]}$, then $x \succ y$ by Suppes-Sen. If $x_{[1]} \leq y_{[n]}$, consider the real numbers $a, b, c, d, e \in \mathbb{R}$ such that $y_{[1]} < a < b < c < x_{[1]} \leq y_{[n]} < d$.

Let $k \in \mathbb{N}$ be the number in the statement of Step 1, which is the same as the k in Step 2. By Step 1, $(c)_{kn} \succsim \big((d)_1, (b)_{kn-1}\big)$. By Step 2, $\big((d)_1, (b)_{kn-1}\big) \succsim \big((d)_{k(n-1)}, (a)_k\big)$. Hence, by transitivity, $(c)_{kn} \succsim \big((d)_{k(n-1)}, (a)_k\big)$. By Replication invariance, this means that $(c)_n \succsim \big((d)_{(n-1)}, (a)_1\big)$. Given that $c < x_{[1]}$, Suppes-Sen implies that $x \succ (c)_n$. Similarly, given that $y_{[1]} < a$ and $y_{[n]} < d$, Suppes-Sen implies that $\big((d)_{(n-1)}, (a)_1\big) \succ y$. By transitivity, $x \succ y$.

Proof of Theorem 1

It is straightforward to check that Maximin SWOs satisfy Suppes-Sen, Consistency, Continuity, Limited sacrifice for the rich future and Limited sacrifice for the long future.

Assume that an SWO $\succsim$ on X satisfies Suppes-Sen, Pigou-Dalton, Consistency, Continuity and Limited sacrifice for the rich future (resp. Limited sacrifice for the long future). By Proposition 1, for all $n \in \mathbb{N}$ and $x, y \in \mathbb{R}^n$, $x \succ y$ whenever $x_{[1]} > y_{[1]}$.

Thus, consider any $n \in \mathbb{N}$ and $x, y \in \mathbb{R}^n$, and assume without loss of generality that $x_{[1]} \geq y_{[1]}$. If $x_{[1]} > y_{[1]}$ we know that $x \succ y$ as required by the Maximin ordering. The only remaining case is $x_{[1]} = y_{[1]}$. We need to show that in that case $x \sim y$.

[10] Note that $a + m\cdot\varepsilon = b$ and $d - m\cdot\varepsilon = c$.

To do so, let us prove that, for any $z \in \mathbb{R}^n$, $z \sim (z_{[1]})_n$. For any sequence of real numbers $(a_1, a_2, \cdots, a_k, \cdots)$ that converges to $z_{[1]}$ and such that $a_k > z_{[1]}$, we have $(a_k)_n \succ z$ because $a_k > z_{[1]}$. Hence, by Continuity $(z_{[1]})_n \succsim z$. Similarly, for any sequence of real numbers $(b_1, b_2, \cdots, b_k, \cdots)$ that converges to $z_{[1]}$ and such that $b_k < z_{[1]}$, we have $z \succ (b_k)_n$ because $b_k < z_{[1]}$. Hence, by continuity $z \succsim (z_{[1]})_n$. Therefore $z \sim (z_{[1]})_n$.

So, when $x_{[1]} = y_{[1]} = a$, $x \sim (a)_n$ and $y \sim (a)_n$. By transitivity, $x \sim y$.

Proof of Theorem 2

Extended continuity: A Lemma

Let us first introduce the following property proposed by Blackorby, Bossert and Donaldson (2001).

Extended continuity. For all $k, \ell \in \mathbb{N}$ and all $x \in \mathbb{R}^k$, the sets $\{y \in \mathbb{R}^\ell \mid y \succsim x\}$ and $\{y \in \mathbb{R}^\ell \mid x \succsim y\}$ are closed in $\mathbb{R}^\ell$.

Note that Extended continuity implies Continuity. The next lemma proves that Extended continuity is implied by Continuity when Critical level holds.

Lemma 2 *If an SWO $\succsim$ on X satisfies Continuity and Critical level then it satisfies Extended continuity.*

Proof. Consider any $k, \ell \in \mathbb{N}$ and all $x \in \mathbb{R}^k$. By repeated applications of Critical level, there exists $z_x \in \mathbb{R}^\ell$ such that $x \sim z_x$. By transitivity, $\{y \in \mathbb{R}^\ell \mid y \succsim x\} = \{y \in \mathbb{R}^\ell \mid y \succsim z_x\}$ and $\{y \in \mathbb{R}^\ell \mid x \succsim y\} = \{y \in \mathbb{R}^\ell \mid z_x \succsim y\}$. By continuity, the sets $\{y \in \mathbb{R}^\ell \mid y \succsim z_x\}$ and $\{y \in \mathbb{R}^\ell \mid z_x \succsim y\}$ are closed in $\mathbb{R}^\ell$, and therefore so are the sets $\{y \in \mathbb{R}^\ell \mid y \succsim x\}$ and $\{y \in \mathbb{R}^\ell \mid x \succsim y\}$. ■

A general class of Maximin criteria

Assuming that a Maximin SWO also satisfies Critical level, we obtain the following Proposition.

Proposition 6 *If a Maximin SWO $\succsim$ on X satisfies Critical level then there exists a function $V : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$, which is non-decreasing in its first argument (and constant when the second argument is negative), continuous and increasing in its second argument, and such that, for all $x, y \in X$,*

$$x \succsim y \iff V\Big(n(x), x_{[1]}\Big) \geq V\Big(n(y), y_{[1]}\Big).$$

Proof. Given that $\succsim$ is a Maximin SWO, for any $k \in \mathbb{N}$, there exists a representative welfare function $e_k : \mathbb{R}^k \to \mathbb{R}$ satisfying $x \succsim y \iff e_k(x) \geq e_k(y)$ for all $x, y \in \mathbb{R}^k$ and $x \sim (a)_k$ whenever $a = e_k(x)$. Specifically, function e_k is such that $e_k(x) = x_{[1]}$ for all $x \in \mathbb{R}^k$.

By Lemma 2, the SWO $\succsim$ satisfies Extended continuity (because Maximin SWOs satisfy Continuity, and $\succsim$ satisfies Critical-level). As shown by Blackorby, Bossert and Donaldson (2001), given that there exists a representative welfare function for each $k \in \mathbb{N}$

and that $\succsim$ satisfies Extended continuity, there exists a function $V : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$, which is continuous and increasing in its second argument, such that, for all $x, y \in X$,

$$x \succsim y \iff V\big(n(x), e_{n(x)}(x)\big) \geq V\big(n(y), e_{n(y)}(y)\big).$$

Given that $e_{n(x)}(x) = x_{[1]}$ for all $x \in X$ and that V is increasing in its second argument, this implies that for all $x, y \in X$,

$$x \succsim y \iff V\Big(n(x), x_{[1]}\Big) \geq V\Big(n(y), y_{[1]}\Big) \tag{8}$$

Let us prove that V is not decreasing in its first argument. Consider any $a \in \mathbb{R}$ and $n \in \mathbb{N}$. By Critical level, there exists $b \in \mathbb{R}_{++}$ such that $(a)_n \sim \big((a)_n, b\big)$. If $b > a$, because $\succsim$ is Maximin, $\big((a)_n, b\big) \sim (a)_{n+1}$. If $b \leq a$, because $\succsim$ is Maximin, $\big((a)_n, b\big) \precsim (a)_{n+1}$. So in any case, by transitivity, $(a)_{n+1} \succsim (a)_n$. By Eq. (8), this implies that $V(n+1, a) \geq V(n, a)$ for any any $a \in \mathbb{R}$ and $n \in \mathbb{N}$: V is non-decreasing in its first argument.

Furthermore, if $a \in \mathbb{R}_{--}$, for any $n \in \mathbb{N}$, Critical level implies that there exists $b \in \mathbb{R}_{++}$ such that $(a)_n \sim \big((a)_n, b\big)$. But given that $b \in \mathbb{R}_{++}$ and $a \in \mathbb{R}_{--}$, because $\succsim$ is Maximin, $\big((a)_n, b\big) \sim (a)_{n+1}$. Hence $(a)_n \sim (a)_{n+1}$. By Eq. (8), this implies that $V(n+1, a) = V(n, a)$ for any any $a \in \mathbb{R}_{--}$ and $n \in \mathbb{N}$: V is constant in its first argument when the second argument is negative. ■

Proof of Theorem 2

It is straightforward to check that Population-adjusted maximin SWOs are Maximin SWOs that satisfy Critical-level and Scale invariance.

Assume that a Maximin SWO $\succsim$ on X satisfies Critical level and Scale invariance. By Prop. 6, we know that there exists a function function $V : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$, which is non-decreasing in its first argument (and constant when the second argument is negative), continuous and increasing in its second argument, and such that, for all $x, y \in X$,

$$x \succsim y \iff \min_{i \in I_{n(x)}} V\big(n(x), x_i\big) \geq \min_{i \in I_{n(y)}} V\big(n(y), y_i\big). \tag{9}$$

Without loss of generality, let normalize function V so that $V(1, a) = a$ for all $a \in \mathbb{R}$ (this is possible given that $V(1, .)$ is increasing and continuous). Given that V is constant when its second argument is negative, we thus have $V(n, a) = a$ for all $n \in \mathbb{N}$ and all $a \in \mathbb{R}_{--}$.

On the other hand, by repeated application of Critical level and because $\succsim$ is Maximin, for any $k \in \mathbb{N}$ there exists $a_k \in \mathbb{R}_+$ such that $(1)_1 \sim (a_k)_k$. By Scale invariance, for any $b \in \mathbb{R}_+$, we have $(b)_k \sim \left(\frac{b}{a_k}\right)_1$ (indeed, $b = \lambda a_k$, with $\lambda = b/a_k > 0$). For any $k \in \mathbb{N}$, denote $\kappa_k = 1/a_k > 0$ (with $a_1 = 1$ so that $\kappa_1 = 1$). By Eq. (9) and the normalization of the V function, we obtain $V(k, a) = \kappa_k \cdot a$ for any $a \in \mathbb{R}_+$. Given that V is non-decreasing in its first argument, the sequence $(\kappa_k)_{k \in \mathbb{N}}$ must be a non-decreasing sequence.

Proof of Proposition 2

Assume that an SWO $\succsim$ on X is a Population-adjusted maximin SWO with population weights $(\kappa_k)_{k \in \mathbb{N}}$.

Let us first show that $\succsim$ avoids the sadistic conclusion. Take any $x \in X$, $k, \ell \in \mathbb{N}$, $a \in \mathbb{R}_{++}$ and $b \in \mathbb{R}_{--}$. We need to show that $\big(x,(a)_k\big) \succsim \big(x,(b)_m\big)$. There are two cases. If $x_{[1]} \leq b < 0$, then, by definition of Population-adjusted maximin SWOs, $\big(x,(a)_k\big) \sim (x_{[1]})_{n(x)+k}$, $(x_{[1]})_{n(x)+m} \sim \big(x,(b)_m\big)$ and $(x_{[1]})_{n(x)+k} \sim (x_{[1]})_{n(x)+m}$, so that, by transitivity, $\big(x,(a)_k\big) \sim \big(x,(b)_m\big)$. On the other hand, if $x_{[1]} > b$ and letting $c = \min\{x_{[1]}, a\} > b$, by definition of Population-adjusted maximin SWOs, $\big(x,(a)_k\big) \sim (c)_{n(x)+k}$, $(b)_{n(x)+m} \sim \big(x,(b)_m\big)$ and $(c)_{n(x)+k} \succ (b)_{n(x)+m}$, so that, by transitivity, $\big(x,(a)_k\big) \succ \big(x,(b)_m\big)$.

Let us then assume that $\succsim$ also satisfies avoidance of the repugnant conclusion. Hence, there exist $k \in \mathbb{N}$, $a \in \mathbb{R}_{++}$ and $b \in [0,a]$ such that, for all $m \geq k$, $(a)_k \succsim (b)_m$. By definition of Population-adjusted maximin SWOs, this implies that $\kappa_m \cdot b \leq \kappa_k \cdot a$, that is $\kappa_m \leq \kappa_k \cdot \frac{a}{b}$, for all $m \geq k$. This is the definition of the sequence $(\kappa_k)_{k\in\mathbb{N}}$ being bounded. It is straightforward to see that, reciprocally, if sequence $(\kappa_k)_{k\in\mathbb{N}}$ is bounded, then there exist $k \in \mathbb{N}$, $a \in \mathbb{R}_{++}$ and $b \in [0,a]$ such that, for all $m \geq k$, $(a)_k \succsim (b)_m$.

Let us now show that $\succsim$ cannot satisfy the Mere addition principle. Indeed, there necessarily exist $a > b > 0$ such that $a > (1+\kappa_2)b$. Thus, by definition of Population-adjusted maximin SWOs, $\big((a)_1,(b)_1\big) \sim (b)_2$ and $(a)_1 \succ (b)_2$ so that by transitivity $(a)_1 \succ \big((a)_1,(b)_1\big)$. This is a violation of the Mere addition principle.

Lastly, let us now show that $\succsim$ cannot satisfy the Negative mere addition principle. Indeed, by definition of Population-adjusted maximin SWOs, $(b)_1 \sim (b)_2$ for any $b \in \mathbb{R}_{--}$. This is a violation of the Negative mere addition principle.

Proof of Proposition 5

Egalitarian case

Let us consider the following function of $r \in \mathbb{R}_+$:

$$V_\theta^E(r) = \tfrac{N^\theta(1+r)^\theta}{1-\varepsilon} \times \left[\left(\tfrac{1+\tilde{g}}{1+\tilde{g}+r} \right)^{1-\varepsilon} \tilde{\omega}^{1-\varepsilon} - 1 \right].$$

Function V_θ^E has the same value as function U_θ^E provided that the egalitarian EDEW is larger than 0. Given that $EDEW^E(0) = u(\tilde{\omega}) > 0$ (because $tildew > 1$), there exists values of r where V_θ^E and U_θ^E are the same and only such values can be optima (when $EDEW^E(r) \leq 0$, social welfare is lower than when $r = 0$). Hence solving problem (5) is equivalent to maximizing V_θ^E.

Now, we have:

$$\begin{aligned}
\tfrac{\partial V_\theta^E}{\partial r} &= \tfrac{N^\theta(1+r)^{\theta-1}}{1-\varepsilon} \times \left[\theta \left(\tfrac{1+\tilde{g}}{1+\tilde{g}+r} \right)^{1-\varepsilon} \tilde{\omega}^{1-\varepsilon} - \theta \right] + \tfrac{N^\theta(1+r)^\theta}{1-\varepsilon} \times \left(\tfrac{\varepsilon-1}{1+\tilde{g}+r} \right) \left(\tfrac{1+\tilde{g}}{1+\tilde{g}+r} \right)^{1-\varepsilon} \tilde{\omega}^{1-\varepsilon} \\
&= \tfrac{N^\theta(1+r)^{\theta-1}}{1-\varepsilon} \times \left[\left(\theta + (\varepsilon-1)\tfrac{1+r}{1+\tilde{g}+r} \right) \left(\tfrac{1+\tilde{g}}{1+\tilde{g}+r} \right)^{1-\varepsilon} \tilde{\omega}^{1-\varepsilon} - \theta \right]
\end{aligned}$$

When $\varepsilon > 1$, the function $H_\theta^E(r) = \left(\theta + (\varepsilon-1)\tfrac{1+r}{1+\tilde{g}+r} \right) \left(\tfrac{1+\tilde{g}}{1+\tilde{g}+r} \right)^{1-\varepsilon} \tilde{\omega}^{1-\varepsilon} - \theta$ is increasing in r and tends to $+\infty$ when r tends to $+\infty$. Thus:

1. if $H_\theta^E(0) < 0$, then H_θ^E is first negative and then positive, so that $\frac{\partial V_\theta^E}{\partial r}$ is first positive and then negative. Function V_θ^E is first increasing and then decreasing and thus admits a unique maximum $r_\theta^{E\star} > 0$, which such that $H_\theta^E(r_\theta^{E\star}) = 0$.

2. if $H_\theta^E(0) \geq 0$, then H_θ^E is always positive (except perhaps at $r = 0$), so that $\frac{\partial V_\theta^E}{\partial r}$ is negative. Function V_θ^E is always decreasing so that its maximum is reached at $r_\theta^{E\star} = 0$.

To sum up, the egalitarian optimal population growth level $r_\theta^{E\star}$ is strictly positive provided $\tilde{\omega} > \left[1 + \frac{\varepsilon-1}{\theta\,(1+\tilde{g})}\right]^{\frac{1}{\varepsilon-1}}$.[11]

Utilitarian case

Using the results from Section 4, we can rewrite utilitarian objective function in Eq. (4) when $\theta > 1$ as follows:

$$U_\theta^U(r) = V_\theta^p\left((1+r)N, \left(1 + r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\right)^{\varepsilon}(1+r)^{-1}\tfrac{\tilde{\omega}^{1-\varepsilon}}{1-\varepsilon} - \tfrac{1}{1-\varepsilon}\right).$$

Let us consider the following function of $r \in \mathbb{R}_+$:

$$V_\theta^U(r) = \tfrac{N^\theta(1+r)^\theta}{1-\varepsilon} \times \left[\left(1 + r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\right)^{\varepsilon}(1+r)^{-1}\,\tilde{\omega}^{1-\varepsilon} - 1\right].$$

We can use a line of arguments similar to the one developed for the egalitarian case to show that solving problem (4) is equivalent to maximizing V_θ^U when $\tilde{\omega} > 1$.

Now, we have:

$$\begin{aligned}
\tfrac{\partial V_\theta^U}{\partial r} &= \tfrac{N^\theta(1+r)^{\theta-1}}{1-\varepsilon} \times \left[\theta\left(1 + r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\right)^{\varepsilon}(1+r)^{-1}\,\tilde{\omega}^{1-\varepsilon} - \theta\right] \\
&\quad + \tfrac{N^\theta(1+r)^\theta}{1-\varepsilon} \times \varepsilon(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\left(1 + r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\right)^{\varepsilon-1}(1+r)^{-1}\,\tilde{\omega}^{1-\varepsilon} \\
&\quad - \tfrac{N^\theta(1+r)^{\theta-1}}{1-\varepsilon} \times \left(1 + r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\right)^{\varepsilon}(1+r)^{-1}\,\tilde{\omega}^{1-\varepsilon} \\
&= \tfrac{N^\theta(1+r)^{\theta-1}}{1-\varepsilon} \times H_\theta^U(r),
\end{aligned}$$

where

$$H_\theta^U(r) = \left(\theta - 1 + \varepsilon(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\tfrac{1+r}{1+r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}}\right)\left(1 + r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\right)^{\varepsilon}(1+r)^{-1}\,\tilde{\omega}^{1-\varepsilon} - \theta.$$

When $\varepsilon > 1$ and $\ln(1+\tilde{g}) \leq 1$, then $\varepsilon(1+\tilde{g})^{\frac{1}{\varepsilon}-1} \geq 1$ and the functions $x \to \varepsilon(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\frac{1+r}{1+r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}}$ and $x \to \left(1 + r(1+\tilde{g})^{\frac{1}{\varepsilon}-1}\right)^{\varepsilon}(1+r)^{-1}$ are increasing. Function H_θ^U is increasing in r and tends to $+\infty$ when r tends to $+\infty$. Thus:

1. if $H_\theta^U(0) < 0$, then H_θ^U is first negative and then positive, so that $\frac{\partial V_\theta^U}{\partial r}$ is first positive and then negative. Function V_θ^U is first increasing and then decreasing and thus admits a unique maximum $r_\theta^{U\star} > 0$, which such that $H_\theta^U(r_\theta^{U\star}) = 0$.

[11] This condition corresponds to $H_\theta^E(0) < 0$.

5. Maximin with population priority $\succsim_{MPP}$: for all $x, y \in X$, if $n(x) > n(y)$ then $x \succ y$, if $n(y) > n(x)$ then $y \succ x$, and if $n(x) = n(y)$ then $x \succsim_{MPP} y \iff x_{[1]} \geq y_{[1]}$;

6. Power population maximin $\succsim_{PPM}$: For all $x \in X$, let $\Gamma : X \to \mathbb{R}$ be such that $\Gamma(x) = x_{[1]}$ if $x_{[1]} \leq 0$ and $\Gamma(x) = (1 + x_{[1]})^{n(x)} - 1$ if $x_{[1]} > 0$. For all $x, y \in X$, $x \succsim_{PPM} y \iff \Gamma(x) \geq \Gamma(y)$.

The Negative utilitarian SWO $\succsim_{NU}$ satisfies Consistency, Continuity, Limited sacrifice for the rich future and Limited sacrifice for the long future, Critical level and Ratio-scale invariance but not Suppes-Sen. The Partial utilitarian SWO $\succsim_{PU}$ satisfies Suppes-Sen, Continuity, Limited sacrifice for the rich future, Limited sacrifice for the long future (for $k = 3$), Critical level and Ratio-scale invariance but not Consistency. The leximin SWO $\succsim_{L}$ satisfies Suppes-Sen, Consistency, Limited sacrifice for the rich future, Limited sacrifice for the long future, Critical level and Ratio-scale invariance but not Continuity. The utilitarian SWO $\succsim_{U}$ satisfies Suppes-Sen, Consistency, Continuity, Critical level and Ratio-scale invariance but neither Limited sacrifice for the rich future nor Limited sacrifice for the long future. The Maximin with population priority SWO $\succsim_{MPP}$ satisfies Suppes-Sen, Consistency, Continuity, Limited sacrifice for the rich future, Limited sacrifice for the long future and Ratio-scale invariance but not Critical level. The Power population maximin SWO $\succsim_{PPM}$ satisfies Suppes-Sen, Consistency, Continuity, Limited sacrifice for the rich future, Limited sacrifice for the long future and Critical level but not Ratio-scale invariance.

Population-adjusted egalitarianism

Supplementary materials

S.A Weak limited sacrifice for the rich future.

The next principle states that if the (single-person of the) current generation is poorer and all future generations equally well-off, then there is a maximal amount of sacrifice we can require of this generation provided all future generations gain enough. Conversely, there is a small loss that is tolerable for all the well-off future generations, no matter how numerous they are, provided the current poorer generation gains enough.

Weak limited sacrifice for the rich future. There exist $\tilde{\alpha}, \tilde{\beta} \in \mathbb{R}_{++}$ such that $\tilde{\alpha} > \tilde{\beta}$ and, for all $n \in \mathbb{N}$, if $a, b, c, d \in \mathbb{R}$ are such that $b \leq c, b - a \geq \tilde{\alpha}$ and $\tilde{\beta} \geq d - c$, then $\big((b)_1, (c)_n\big) \succsim \big((a)_1, (d)_n\big)$.

Weak limited sacrifice for the rich future is related to a principle named Weak non-aggregation by Fleurbaey and Tungodden (2010). The next Proposition states if this principle is satisfied together with Ratio-Scale invariance then we obtain the principle of Limited sacrifice for the rich future used in the main text.

Proposition S.A.1 *Consider an SWO $\succsim$ on X. If $\succsim$ satisfies Weak limited sacrifice for the rich future and Ratio-scale invariance then it satisfies Limited sacrifice for the rich future.*

Proof. Consider any $\alpha \in \mathbb{R}_{++}$. Let $\tilde{\alpha}$ be the number in the statement of Weak limited sacrifice for the rich future, $\lambda = \frac{\tilde{\alpha}}{\alpha} > 0$ and $\beta = \tilde{\beta}/\lambda$ where $\tilde{\beta} \in \mathbb{R}_{++}$ be the number in the statement of Weak limited sacrifice for the rich future (hence $\alpha > \beta > 0$).

Consider any $n \in \mathbb{N}$ and any $a, b, c, d \in \mathbb{R}$ are such that $b \leq c, b - a \geq \alpha$ and $\beta \geq d - c$. Thus $\lambda\, a, \lambda\, b, \lambda\, c, \lambda\, d$ are real numbers such that $\lambda\, b \leq \lambda\, c$, $\lambda\, b - \lambda\, a \geq \lambda\, \alpha = \tilde{\alpha}$ and $\tilde{\beta} = \lambda\, \beta \geq \lambda\, d - \lambda\, c$. By Weak limited sacrifice for the rich future, this implies that $\big((\lambda b)_1, (\lambda c)_n\big) \succsim \big((\lambda a)_1, (\lambda d)_n\big)$. And by Ratio-scale invariance, this yields $\big((b)_1, (c)_n\big) \succsim \big((a)_1, (d)_n\big)$. ■

S.B Population-adjusted leximin social welfare orderings

In this section, I show how the axiomatic analysis of Section 3 can be extended to the case of Leximin SWOs. A difficulty is that leximin SWOs are nor continuous, so that we

cannot define an EDEW and combine it with population size like in Blackorby, Bossert and Donaldson (2001). To compare populations with different sizes, I thus use a different approach.

A key property of leximin SWOs is that they satisfy a strong notion of Consistency that requires that if the situation is strictly socially better for one subpopulation it is also socially better for the whole population. This can be expressed as an independence axiom.

Weak independence For all $n \in \mathbb{N}$, all $x, y \in \mathbb{R}^n$ and all $a, b \in \mathbb{R}$, $(a, x) \succsim (a, y)$ if and only if $(b, x) \succsim (b, y)$.

Furthermore, for all $x, y \in X$ and all $a, b \in \mathbb{R}_-$, $(a, x) \succsim (a, y)$ if and only if $(b, x) \succsim (b, y)$.

This a weak existence independence principle because independence holds only for allocations with the same population size or when independence is with respect to individuals with non-positive level of well-being.

Proposition S.B.1 *Consider an SWO $\succsim$ on X.*

1. *If $\succsim$ satisfies Suppes-Sen, Weak independence and Limited sacrifice for the rich future, then it is a Leximin SWO.*
2. *If $\succsim$ satisfies Suppes-Sen, Weak independence and Limited sacrifice for the long future, then it is a Leximin SWO.*

Proof. Let us first show that Weak independence implies Consistency. Consider any $n, m \in \mathbb{N}$, any $x, y \in \mathbb{R}^n$ and any $x', y' \in \mathbb{R}^m$. Assume that $x \succsim y$ and $x' \succsim y'$. By repeated applications of Weak independence, $(x, x') \succsim (x, y')$ and $(x, y') \succsim (y, y')$. By transitivity, $(x, x') \succsim (y, y')$. The same line of reasoning implies that if $x \succ y$ and $x' \succ y'$ then $(x, x') \succ (y, y')$.

Now, by Prop. 1, given the axioms in the two statements, we know that for all $n \in \mathbb{N}$ and $x, y \in \mathbb{R}^n$, if $x_{[1]} > y_{[1]}$ then $x \succ y$.

Consider any $n \in \mathbb{N}$ and any $x, y \in \mathbb{R}^n$. By Suppes-Sen, if $x_{[\,]} = y_{[\,]}$, then $x \sim y$. If $x_{[\,]} \neq y_{[\,]}$, there must exist $R \in I_n$ such that $x_{[r]} = y_{[r]}$ for all $r \in I_{R-1}$ and either $x_{[R]} > y_{[R]}$ or $y_{[R]} > x_{[R]}$. Consider without loss of generality the case $x_{[R]} > y_{[R]}$ and let $a, b \in \mathbb{R}$ be such that $x_{[R]} > a > y_{[R]}$ and $b > \max\{x_{[R]}, y_{[n]}\}$. Define $z \in \mathbb{R}^{R-1}$, $\tilde{x}, \tilde{y} \in \mathbb{R}^{n-R+1}$ in the following way:

- For all $i \in I_{R-1}$, $z_i = x_{[i]} = y_{[i]}$;
- For all $j \in I_{n-R+1}$, $\tilde{x}_j = x_{[R]}$;
- $\tilde{y}_1 = a$ and for all $j \in \{2, \cdots, n-R+1\}$, $\tilde{y}_j = b$.

By Suppes-Sen, $x \succsim (z, \tilde{x})$ and $(z, \tilde{y}) \succsim y$. Let $\tilde{x}' = \big((b)_{R-1}, \tilde{x}\big)$ and $\tilde{y}' = \big((b)_{R-1}, \tilde{y}\big)$, so that $\tilde{x}'_{[1]} = x_{[R]} > a = \tilde{y}'_{[1]}$. Hence, $\tilde{x}' \succ \tilde{y}'$. By repeated applications of Weak independence this implies $(z, \tilde{x}) \succ (z, \tilde{y})$ and by transitivity that $x \succ y$. ■

To apply Leximin SWOs to variable populations comparisons, I propose to accept the following two principles. The first principles ensures that there exists a trade-off between average welfare and population size.

Sensible trade-off. For any $n \in \mathbb{N}$, there exist $a, b \in \mathbb{R}_{++}$ such that $a \geq b$ and $(a)_n \sim (b)_{n+1}$.

A stronger property than the Critical level used in the main text requires all critical levels to be equal to the utility level representing neutrality. This is what Blackorby, Bossert and Donaldson (2005) name the Zero critical level principle. As they argue, this may not be a compelling axiom in general, but here we restrict it to 'bad lives' (below neutrality). In that case, the addition of people with non-negative well-being seems to be an improvement. Combined with Suppes-Sen, the axiom also ensures that the Weak mere addition principle (defined in the main text) is satisfied.

Zero critical level for bad lives. For any $x \in X_-$, $\big(x, (0)_1\big) \sim x$.

I also propose a equivalence condition that resembles the Suppes-Sen principle but can be applied to populations with different sizes. To do so, we need to take into account population size when assessing the contributive value of lives.

Population-adjusted Suppes-Sen equivalence. For all $x, y \in X$, if $n(x) > n(y)$, $(x_{[\ell]})_{n(x)} \sim (y_{[\ell]})_{n(y)}$ for all $\ell \leq n(y)$, $x_{[n(y)]} \geq 0$ and $x_{[n(x)]} = x_{[n(y)]}$, then $x \sim y$.

To define Population-adjusted leximin SWOs, let me introduce a new piece of notation. For an allocation $x \in \mathbb{R}^n$, a sequence $(\kappa_k)_{k\in\mathbb{N}} \in \mathbb{R}_+^{\mathbb{N}}$ and a positive integer $m \in \mathbb{N}$ such that $m \geq n$, we denote $x^{\kappa,m}$ the allocation such that:

- If $m = n$, then $x_i^{\kappa,m} = \Pi_\kappa(n, x)$ for all $i \in I_n$;[2]

  If $m > n$ then $x_i^{\kappa,m} = \Pi_\kappa(n, x_{[i]})$ for all $i \in I_n$ and $x_j^{\kappa,m} = \Pi_\kappa\Big(n, \max\{x_{[n]}, 0\}\Big)$ for all $j \in \{n+1, \cdots, m\}$.

Definition S.B.1 *An SWO $\succsim$ on X is a Population-adjusted leximin SWO if and only if there exists a non-decreasing sequence $(\kappa_k)_{k\in\mathbb{N}}$ such that $\kappa_1 = 1$ and for all $x, y \in X$ with $n(x) \geq n(y)$,*

$$x \succsim y \iff x^{\kappa,n(x)} \succsim_L^{n(x)} y^{\kappa,n(x)}$$

and

$$y \succsim x \iff y^{\kappa,n(x)} \succsim_L^{n(x)} x^{\kappa,n(x)}.$$

The next Theorem is a characterization of Population-adjusted leximin SWOs.

Theorem S.B.1 *Consider an SWO $\succsim$ on X.*

1. *$\succsim$ satisfies Suppes-Sen, Limited sacrifice for the rich future, Ratio-scale invariance, Weak independence, Sensible trade-off, Zero critical level for bad lives and Population-adjusted Suppes-Sen equivalence if and only if it is a Population-adjusted leximin SWO.*

---

[2] Recall that function $\Pi_\kappa : \mathbb{N} \times \mathbb{R}$ is defined for all $(n, e) \in X \times \mathbb{R}$ by:

$$\Pi\kappa(\text{n,e}) = \begin{cases} e\& \text{ if } e \leq 0; \\ \kappa n \cdot e\& \text{ if } e > 0. \end{cases}$$

2. *$\succsim$ satisfies Suppes-Sen, Limited sacrifice for the long future, Ratio-scale invariance, Weak independence, Sensible trade-off, Zero critical level for bad lives and Population-adjusted Suppes-Sen equivalence if and only if it is a Population-adjusted leximin SWO.*

Proof. It is straightforward to check that Population-adjusted leximin SWOs satisfy all the principles in the two statements.

Assume that an SWO $\succsim$ on X satisfies the axioms in principles in the two statements. Let us show that it is a Population-adjusted leximin SWO.

Step 1: $\succsim$ is a Leximin SWO.

By Prop. S.B.1, given the principles in the two statements, we know that $\succsim$ is a Leximin SWO.

Step 2: For any $x \in X_-$, for any $m > n(x)$, $x \sim x^{\hat{\kappa},m}$ for any sequence $(\hat{\kappa}_k)_{k\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}_+$.

By repeated applications of Zero critical level for bad lives $x \sim \big(x, (0)_{m-n(x)}\big)$. By definition, $\big(x, (0)_{m-n(x)}\big) = x^{\hat{\kappa},m}$ for any sequence $(\hat{\kappa}_k)_{k\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}_+$. Remark that this implies that $(0)_m \sim (0)_n$ for all $m, n \in \mathbb{N}$.

Step 3: There exists a non-decreasing sequence $(\kappa_k)_{k\in\mathbb{N}}$ such that $\kappa_1 = 1$ and for any $a, b \in R_{++}$ and $n, m \in \mathbb{N}$ if $\kappa_n \cdot a = \kappa_n \cdot b$ then $(a)_n \sim (b)_m$.

By Sensible trade-off, for any $n \in \mathbb{N}$ there exists $a, b \in \mathbb{R}_{++}$ such that $a \geq b$ and $(a)_n \sim (b)_{n+1}$. Denote $\mu_n = \frac{a}{b} \geq 1$ so that $(\mu_n b)_n \sim (b)_{n+1}$. By Scale invariance, for any $c \in \mathbb{R}_{++}$, there exists $d \in \mathbb{R}_{++}$ such that $d = \frac{c}{b}a = \mu_n c$ such that $(d)_n \sim (c)_{n+1}$ (indeed, $c = \lambda b$ and $d = \lambda a$ with $\lambda = \frac{c}{b}$). This implies that there exists $\mu_n \geq 1$ such that, for $c \in \mathbb{R}_{++}$, $(\mu_n c)_n \sim (c)_{n+1}$.

By transitivity and repeated application of the above procedure, for any $a \in \mathbb{R}_{++}$ and $n \in \mathbb{N}$ we have $(a)_n \sim (\kappa_n a)_1$, where $\kappa_1 = 1$ and for any $n > 1$, $\kappa_n = \left(\prod_{k=1}^{n-1} \mu_k\right)$. Remark that $\kappa_{n+1} \geq \kappa_n$ for all $n \in \mathbb{N}$ given that $\mu_k \geq 1$ for all $k \in \mathbb{N}$. By transitivity, we obtain that for any $a, b \in \mathbb{R}_{++}$ and any $n, k \in \mathbb{N}$, if $\kappa_n \cdot a = \kappa_n \cdot b$ then $(a)_n \sim (b)_m$.

Step 4: There exists a non-decreasing sequence $(\kappa_k)_{k\in\mathbb{N}}$ such that $\kappa_1 = 1$ and, for all $n, m \in \mathbb{N}$ with $m \geq n$, for all $x \in \mathbb{R}^n$, $x \sim x^{\kappa^m,m}$, where $\kappa^m = (\kappa_1/\kappa_m, \cdots, \kappa_{m-1}/\kappa_m, 1, 1, \cdots)$.

By Step 2, we know that this is true for all $x \in X_-$. Thus consider any $n, m \in \mathbb{N}$ with $m \geq n$ and any $x \in \mathbb{R}^n \setminus \mathbb{R}^n_{--}$. Let $R \in I_n$ be the integer such that $x_{[r]} < 0$ for all $r < R$ and $x_{[r]} \geq 0$ for all $r \geq R$. Let us assume that $R > 1$ (the proof for $R = 1$ is similar but does not need the complication of dealing with negative welfare levels).

For $n = m$, the statement is true as $x \sim x^{\kappa^m,m}$. Assume that $m > n$ and let $(\kappa_k)_{k\in\mathbb{N}}$ be the sequence in Step 3. Define $z \in \mathbb{R}^{R-1}$, $\tilde{x} \in \mathbb{R}^{n-R+1}$ and $\tilde{y} \in \mathbb{R}^{m-R+1}$ in the following way:

- For all $i \in I_{R-1}$, $z_i = x_{[i]}$;
- For all $j \in I_{n-R+1}$, $\tilde{x}_j = x_{[R-1+j]}$ and $\tilde{y}_j = \frac{\kappa_n}{\kappa_m} x_{[R-1+j]}$;
- For all $j \in \{n - R + 2, \cdots, m - R + 1\}$, $\tilde{y}_j = \frac{\kappa_n}{\kappa_m} x_{[n]}$.

Consider allocations $\hat{x} = \big((0)_{R-1}, \tilde{x}\big)$ and $\hat{y} = \big((0)_{R-1}, \tilde{y}\big)$. Clearly, for all $r \in I_{R-1}$ $(\hat{x}_r)_n \sim (\hat{y}_r)_m$ because $x_{[r]} = y_{[r]} = 0$ (see Step 2). For all $r \in \{R, \cdots, n\}$, we have $\hat{x}_r = x_{[r]}$ and $\hat{y}_r = \frac{\kappa_n}{\kappa_m} x_{[r]}$. Hence $\kappa_n \hat{x}_r = \kappa_m \hat{y}_r$. Thus, by Step 3, $(\hat{x}_r)_n \sim (\hat{y}_r)_m$ for

all $r \in \{R, \cdots, n\}$. Lastly, $\hat{y}_{[m]} = \hat{y}_{[n]}$. Hence, by Population-adjusted Suppes-Sen equivalence, $\hat{x} \sim \hat{y}$.

By repeated applications of Weak independence, $\left(z, \tilde{x}\right) \sim \left(z, \tilde{y}\right)$. Let $\kappa^m = (\kappa_1/\kappa_m, \cdots, \kappa_{m-1}/\kappa_m,$ By definition, $\left(z, \tilde{y}\right)_{[\,]} = x_{[\,]}^{\kappa^m, m}$ and $\left(z, \tilde{x}\right)_{[\,]} = x_{[\,]}$. By Suppes-Sen and transitivity, this implies $x \sim x^{\kappa^m, m}$.

Step 5: Conclusion. Consider any $x, y \in X$ with $n(x) \geq n(y)$.

By Step 4, $x \sim x^{\kappa^{n(x)}, n(x)}$ and $y \sim y^{\kappa^{n(x)}, n(x)}$ where $(\kappa_k)_{k \in \mathbb{N}}$ is the sequence in the statement of Step 3 and (κ^n, n) is defined in the statement of Step 4. Both $x^{\kappa^{n(x)}, n(x)}$ and $y^{\kappa^{n(x)}, n(x)}$ are allocations in $\mathbb{R}^{n(x)}$ so, by Step 1, $x^{\kappa^{n(x)}, n(x)} \succsim y^{\kappa^{n(x)}, n(x)} \iff x^{\kappa^{n(x)}, n(x)} \succsim_L^{n(x)} y^{\kappa^{n(x)}, n(x)}$ and $y^{\kappa^{n(x)}, n(x)} \succsim x^{\kappa^{n(x)}, n(x)} \iff y^{\kappa^{n(x)}, n(x)} \succsim_L^{n(x)} x^{\kappa^{n(x)}, n(x)}$. Remark that for all $i \in I_{n(x)}$ $x_i^{\kappa, n(x)} = \Pi\left(n(x), x_i^{\kappa^{n(x)}, n(x)}\right)$ and $y_i^{\kappa, n(x)} = \Pi\left(n(x), y_i^{\kappa^{n(x)}, n(x)}\right)$. Given $\Pi(n(x), .)$ is an increasing function and that Leximin orderings are invariant with respect to transformations of utility by a common increasing function we obtain that $x^{\kappa^{n(x)}, n(x)} \succsim y^{\kappa^{n(x)}, n(x)} \iff x^{\kappa, n(x)} \succsim_L^{n(x)} y^{\kappa, n(x)}$ and $y^{\kappa^{n(x)}, n(x)} \succsim x^{\kappa^{n(x)}, n(x)} \iff y^{\kappa, n(x)} \succsim_L^{n(x)} x^{\kappa, n(x)}$. Then transitivity yields the result. ■

Katie Steele[1]

'International Paretianism' and the Question of 'Feasible' Climate Solutions[2]

Proponents of *International Paretianism (IP)*—the principle that international agreements should not make any state worse-off and should make some at least better off—argue that it is the only feasible approach to reducing the harms of climate change (see, especially, Posner and Weisbach 2010). They draw on some key assumptions regarding the meaning of 'feasibility' and the nature of the Pareto improvements associated with coordinated action on climate change. This chapter challenges these assumptions, in effect weakening the case for IP and allowing for broader thinking about what counts as a 'feasible' climate solution.

[1] Institute for Futures Studies and School of Philosophy, Australian National University, katie.steele@iffs.se. Financial support by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences) is gratefully acknowledged.
[2] Forthcoming in *Philosophy and Climate Change*, Budolfson, M., McPherson, T. and Plunkett, D (eds.) Oxford University Press.

1 Introduction

In view of the considerable suffering that climate change threatens in the medium- and long-term future, the global response to this problem has been woefully inadequate. This very concerning state of affairs has lead many to reflect on the nature of the problem and the prospects for solving it. Accordingly, the notion of *political feasibility* has become prominent in debate about who can be counted on to bear the costs of climate change—the costs of mitigation and adaptation. While it is all well and good arguing over what are better and worse distributions of these costs, actual progress can only be achieved through action, which is constrained by what is feasible. Or so the rhetoric might go. Indeed, it is hard to deny that, ultimately, it is important that we actually advance on the status quo—that climate-related suffering is reduced by *some amount*, the more the better, even if this falls far short of a morally good state of affairs.

A specific principle or maxim has been pitched as the answer to the call for 'feasible' climate solutions—*International Paretianism* (IP)—which basically recommends framing international negotiations on problems like climate change in terms of mutual 'national interest'. The idea is that, as far as feasibility goes, any proposal for determining responsibilities for tackling climate change must advance the interests of all states, or at least not frustrate their interests. That is, a genuine climate solution must provide a *Pareto Improvement* with respect to the interests of states: No state is made 'worse-off', and at least one state is 'better off' than they fare under the status quo. One might question whether states really are the key international players in climate negotiations, but let us assume that this feature of international IP is plausible. It is also ambiguous as to what counts as a state's 'self or national interest', but let us assume it is an appropriate aggregate of the welfare of present and future citizens (more about this later). As such, the costs of a mitigation strategy for present citizens can be counterbalanced by the benefits for future citizens, such that there is a net benefit in terms of national interest.

Advocates argue that IP is uniquely placed with respect to political feasibility in the international setting (see, especially, Eric Posner and David Weisbach's 2010 book *Climate Justice*). They claim that the only feasible proposals for significantly reducing the harms of climate change are IP proposals, and thus treaties intended to actually reduce suffering associated with climate change ought to satisfy IP criteria. While IP is not a radically new approach to international agreements, it is nonetheless provocative in climate justice circles.[3]

[3] Note, however, that some philosophers have also claimed that it may be productive, at least in the short term, to focus on climate solutions that amount to Pareto improvements on the status quo; notably Broome (2010); see also Schokkaert and Eyckmans (1998). The focus here is on the work of

Discussions of climate justice have traditionally focussed rather on historical responsibility and duties of reparation, or simply duties of assistance, that are owed by rich, high-emitting states.[4] IP, on the other hand, redirects attention to what is good for rich, high-emitting states (as well as what is good for other states). Thus IP, as an approach to climate change, may seem rather shocking. In spite of this, there is certainly something compelling, hopeful even, about IP, once we focus on political feasibility. IP is not about states recognising their duties and stepping up to make whatever demanding reparations for climate change are morally required of them. It is rather about promoting a climate solution that the relevant actors would have no reason to resist and may have material interest in pursuing. We would be fools to let an overly aspirational treaty prevent easy agreement on a morally inferior, but nonetheless mutually beneficial, one. Or so the thinking might go.

Political feasibility is certainly an important consideration in the debate about climate change. But once we look to the closer details of what feasibility stands for, its relationship with furthering self-interest, let alone the national interest of a state, is not as persuasive as first appears. In order to arrive at a substantial global response to climate change, IP must appeal to a notion of national interest that is rather revisionary given the existing apparent motivations of state actors. Furthermore, closer inspection of the form of the cooperative dilemma associated with climate change reveals that the pursuit of national interest alone may turn out to be self-defeating. This is more worrisome for the feasibility argument for IP than its advocates suggest. Or so this chapter will argue. More generally, the aim here is to shed light on International Paretianism, the climate change predicament, and the meaning and role of 'feasibility' in ethical-political debate. Section 2 begins with the latter conceptual issue. Section 3 turns to the climate predicament itself—the kind of Pareto improvement that may be in the offing, and explores whether the achievement of such a Pareto improvement does in fact depend solely on 'self interest'. Section 4 reflects on the implications of these considerations for the overall feasibility claims made by IP proponents.

Posner and Weisbach because they make stronger, more definitive claims about feasibility in the international setting. Moreover, Broome takes individual persons, rather than nation states, to be the basic agents that are subject to climate-related costs and benefits, and so does not defend International Paretianism *per se*.

[4] Typically, the focus is on the *costs* of mitigation, assuming that an appropriate global mitigation target has been decided. The question is how to divide up the costs of meeting this target; a variety of moral perspectives lead to the conclusion that the rich high-emitting states should bear the bulk of the costs.

2 The concept of 'political feasibility'

As suggested above, Posner and Weisbach (2010) largely rest their case for an IP-style climate treaty on the fact that only this style of climate treaty is 'politically feasible'. [5] But what is the precise meaning and significance of that term? As suggested above, the reason to focus on feasibility at all stems from an interest in actual social change, in 'getting things done'. So 'feasibility' is a property that is invoked in the interests of actual change. Beyond that, however, the proper usage of the term is not obvious. And yet the plausibility of the IP proponents' claim – the tight relationship between feasibility and Pareto improvements – depends on these details.

An initial concern is that the concept is inherently vague and so the IP proponents' claim will resist any definitive analysis. The very target of the concept – what supposedly has the property of 'feasibility' – seems to vary in general talk and debate, let alone the rules for determining whether/to what extent the property is present. We talk about whether a goal, say, of swimming 50m in less than half a minute, is feasible; about whether change, say, towards a less racist society, is feasible; about whether a program, say, a new teaching curriculum, will feasibly improve students' learning outcomes; about whether it is feasible that I will not procrastinate while doing some task. Some of these examples seem to be about whether actions will realistically be performed (e.g., work without procrastination), and others about whether goals will realistically be achieved (e.g., the fast swim or a less racist society) perhaps in addition to or else conditional on some action being performed (e.g., the improvement of students' learning outcomes given reforms to the teaching curriculum).

The further concern is that there are apparently no set criteria for determining whether the intended target – a goal or action or both – counts as feasible, i.e., how plausible its realisation must be and what is the basis for such an assessment. Perhaps we should accept that 'feasibility' is simply a smokescreen for differing substantial views about what are worthwhile pursuits. In the political context, the question might boil down to old disputes about what are viable prospects for social

[5] A qualification must be made here: Posner and Weisbach (2010) also argue that a climate treaty should be concerned only with advancing mutual self-interest because anything more ambitious in terms of rich states paying would amount to trying to solve other problems of global distributive justice in a climate treaty, which would not be an *efficient* way to solve these other important problems. So they argue for IP on efficiency grounds (in the broader scheme of distributive justice) as well as feasibility grounds. It is not at all obvious that the optimal IP climate solution would in fact yield the amount of climate change mitigation that a benevolent global planner would choose, which is an aspect of the efficiency argument. Later in Section 3.1 some further remarks are made on this point. But it is beyond the scope of this chapter to discuss the efficiency argument in detail. See, however, Frisch (2012), Baer (2013), Jamieson (2013) and Shue (2013), for critiques of Posner and Weisbach's claims regarding the ethical optimality, so to speak, of IP.

change. If IP proponents were simply staking a 'realist' position regarding international relations, say, then here again their claims would seem to resist any searching analysis. Quite simply: those sympathetic to the 'realist' view (roughly that 'might is right') in the international arena will be sympathetic to the IP view, unlike those of more 'idealist' persuasion.

There is something to the above concerns, but they are a bit quick. We will see that analysing the IP proponents' claims about feasibility and climate change is more interesting and fruitful than first meets the eye. To begin with, the growing philosophical literature on political feasibility suggests that the concept has subtleties revealing a need to clarify what legitimate role(s) it may play in debate. This will become apparent in the attempt to pin down a working definition of the term in 2.1 below. Note that while there is some divergence amongst the prominent definitions in the literature (and moreover some suggest that 'feasibility' may play multiple normative roles and so accordingly have multiple meanings) there are nonetheless a number of points of agreement amongst feasibility scholars. These points of agreement will be taken as 'fixed points' in our analysis. Beyond that, the strategy here is to fashion a notion of feasibility that is, as it were, charitable to the IP proponents' position. It should be a useful concept for evaluating climate change proposals and flexible enough to *potentially* vindicate the IP proponents' claim regarding a strong relationship between feasibility and Pareto improvements with respect to climate change. This more substantial issue is introduced in 2.2 and considered in more detail in the remainder of the chapter.

2.1 A working definition of 'political feasibility'

Despite there being a range of views about the finer details of the concept of 'political feasibility', there is reasonable agreement about the basics. In particular, there is broad agreement (also in line with the discussion thus far) on the general function of assessments of feasibility: They bear on directives for action and, roughly speaking, concern the possibility of success. In the spirit of 'ought implies can', what an agent *ought to do* is *somehow sensitive* to what is feasible (see Brennan and Southwood 2007, Gilabert and Lawford-Smith 2012, Lawford-Smith 2013, Southwood 2016, Wiens 2015).[6] Note that this accords with the examples given above of how the term is variously employed in debate. All these examples relate to directives for action, although, as noted, the precise nature of the target seems to vary – the

[6] There are many further subtleties concerning this phrase that I will not explicitly address in this discussion – concerning the way and extent to which 'what one ought to do is sensitive to what is feasible'. Southwood (2016), for instance, discusses different notions of 'ought' that have differing logical relationships with his stated notion of 'feasibility'. The point here is simply that the notion of feasibility should be defined in such a way that it plays *some* guiding role in determining right action.

feasibility assessment applies sometimes to an act, sometimes to a goal or state of affairs to be acted upon, and sometimes to a combination of the two. (We will return to this diversity shortly.)

Given the function just described, it follows that feasibility assessments pertain to a given agential positioning or decision-making context (the more precisely specified the more precise will be the feasibility assessment). All agree that feasibility has to do with whether the target may be plausibly realised, but, as mentioned, there is disagreement about the target. Perhaps the most natural target is an action that an agent might consider performing. This is the line taken by Brennan and Southwood (2007) and Southwood (2016). We can interpret these authors as holding that the 'feasible acts' are akin to what decision theorists might refer to as the 'available acts' for some decision-maker, or in other words, the 'choice set'. It follows that feasibility, understood in this way, is a binary matter: Candidate acts are either feasible (i.e., in the choice set), or not, depending on worldly constraints and whether the agent can 'bring him/herself to perform the act in question'[7]. The motivational constraint here is that the agent must be able to decide, if it follows from their psychological attitudes, upon the act in question, to initiate the act if it is chosen, and also to follow through in performing the act.[8] So the agent must be sufficiently likely to actually go ahead and perform the act, conditional on being psychologically disposed, in some sense, to do so.

Others also endorse a binary notion of feasibility that is a property of acts. There are differences, however, in emphasis, and in the kind of 'acts' that are taken to be at issue. Wiens (2015), for instance, attends much more closely to how motivational and resource constraints together affect what acts/ options in the political domain are sufficiently likely to be realised and are thus 'feasible'. In so doing, the acts he considers are rather more complicated ones involving multiple agents. There is not such an obvious connection between feasibility and a decision-maker's 'choice set' (to use the decision theory language introduced above). Gilabert and Lawford-Smith (2012) also extend the target of feasibility assessments beyond simple acts; they include a goal as part of their account, thus proposing a four-place predicate for feasibility: On their view, feasibility pertains to a particular agent performing a specified act in order to achieve some goal in a given context. Again, some of the acts (paired with goals) they consider are rather complicated ones. (They also introduce more and less minimalist feasibility assessments, presumably to cater for the term having multiple roles in ethical discourse; one is a binary notion and depends on

[7] This is, roughly, the phrase used by Southwood (2016, 11), who attributes it to Estlund (2011). Note that even if probabilities are employed in reasoning about what is feasible, on this account ultimately the judgment must be binary because arguably it does not make sense for an act to be only partially in an agent's choice set.

[8] As per Southwood (2016, section II)

'hard' logical-metaphysical constraints, while the latter is a graded notion and depends on further 'soft' empirical details.[9])

One might diagnose the situation as follows. There is disagreement about what is the most useful notion of 'feasibility' when it comes to directives for action. Specifically, should the target be the decision theoretic 'acts' that would, if feasible, be the basic items of a decision-maker's choice set? (The possible consequences of these acts would then be another, further matter.) Or should the target rather be more complicated activities—complex acts that may take years to carry out, say, and are perhaps also associated with a goal? The question is important because it may not always be fruitful to think of complex acts as options that a decision maker can, depending on a binary feasibility assessment, simply choose to perform. At least some kinds of complex acts seem to call for a notion of feasibility that comes in degrees,[10] and accordingly, are not best conceived as the choice options available to a decision maker.

To avoid ambiguity, it helps to introduce a new term for the kinds of 'complex acts' referred to above: Let us call them *projects* or *multi-stage plans*. This seems to best characterise the target for feasibility assessments in the climate debate. (The account developed here need not be the only useful notion of feasibility.) Climate treaties or proposals are more like projects or multi-stage plans than simple acts that a decision maker may or may not choose to perform. Indeed, climate treaties not only concern a course of action that spans some considerable length of time; they also involve more than one agent (with the further complication that these agents are themselves groups). The feasibility (or predicted success) of such a project surely comes in degrees. But given projects are not the sort of thing a single agent may simply decide to perform, does feasibility still concern directives for action and pertain to an agent's deliberations? Yes. For starters, it is hard to make sense of the feasibility of a project or multi-stage plan, complex as this series of acts may be, absent the perspective of a deliberating agent and a specific way of initiating the project (even if these matters of perspective are not always made explicit in our everyday talk). Second, there are good reasons for a decision-maker to assess the feasibility or likelihood of success of a project, were it initiated in a specific way. The most obvious reason is that the decision maker is considering whether it is worth doing the initiation herself. This clearly depends on the likelihood that, were she to take this action, the project would eventually be successfully implemented. That is, the full implementation of the project is one possible consequence of the act of

[9] While Southwood (2016) focuses on different logical roles for 'feasibility', depending on the moral concept at issue, Gilabert and Lawford-Smith seem to focus rather on different definitions of 'feasibility', depending on the moral concept at issue.

[10] As per Gilabert and Lawford-Smith's *graded* notion of feasibility that takes into account 'soft' constraints

initiation, and thus the likelihood of this consequence – the feasibility of the project were it initiated in the given way – is relevant to the agent's deliberations about whether the project ought to be initiated in that way.[11]

It is worth clarifying the project account of feasibility before moving on to the more substantial questions concerning self-interest and Pareto improvements. We see that projects are of concern to a deliberating agent. The difference with basic acts is that, at the point of choice, the agent can merely initiate the project. Subsequent steps in the completion of the project are not directly under the agent's present control; they are the province of other agents or perhaps the 'future selves' of the initiating agent. So projects are rather complex and there is room for a variety of eventualities once they are initiated. One of these eventualities or possible consequences is the full implementation or successful realisation of the project. The likelihood of this outcome is the feasibility of the project.

Note that we refer to projects or multi-stage plans in diverse ways and so many feasibility statements could conceivably refer to projects. (Consider the examples given earlier.) Projects are sometimes described as if they were single actions, albeit complex ones that would take a number of steps to implement over some period of time. Consider the 'act' of me working without procrastinating, or, at the grander scale, the 'act' of Australia reducing its greenhouse gas (GHG) emissions by 26% on 2005 levels by 2030.[12] Note that the latter example of Australia reducing its emissions may equally be regarded a goal. Indeed, projects are often described in the language of goals, for instance, 'the goal of limiting climate change to 2 degrees C rise in average global temperature', or 'the achievement of universal access to high-speed internet'. Moreover, projects may be more or less detailed. The goal of limiting climate change to 2 degrees C may be stated in such a way that the precise means to this end are specified, or else rather left open.[13] But however they are referred to in casual talk, the feasibility of a project, on the account given here, depends on the way it would be initiated by the relevant agent and what the whole project amounts to, or what are the conditions for its success; the more precisely these details are specified, the more precise the feasibility assessment.

A further issue is how we should think about the feasibility of projects involving multiple agents. One may refer loosely to 'group projects' but the idea here is that any project must somehow be initiated, and this concerns the choice of a single

[11] By the same token, the nature and likelihood of the project *not* being successfully realized is also relevant to the agent's deliberations.
[12] As agreed to in Paris in 2015.
[13] Compare the following two projects regarding the mitigation of climate change: i) Reduce greenhouse gas (GHG) emissions by 26% ii) Implement extensive program of carbon-capture-and-storage so as to reduce GHG emissions by 26%. The latter project is obviously more detailed in terms of the steps to be taken, and as such, its feasibility can be more precisely specified.

agent or decision maker.[14] The choice concerns whether to perform the initiating act. Whether the project will succeed then depends, amongst other things, on the *predicted actions of the other agents* who are relevant to the success of the project. In other words, the behaviour of these other agents, from the point of view of the decision maker, are simply aspects of the world, like whether or not it will rain, that bear on the consequences of the initiating act. Note that the decision maker may relate to the wider group in a variety of ways: He/she may be a member of the group who, for the purposes at hand, is regarded the deliberating agent. Or else he/she may be an onlooker to the group who is nonetheless interested in the initiation of the group project. To see the distinction, consider the assessment of a proposed international climate treaty by i) a representative of a participating state and ii) a UN official charged with brokering the deal. These agents may justifiably have differing assessments of the feasibility of the treaty, due to differences in how they would initiate the project.

So feasibility is the probability that a project or multi-stage plan would be successfully realised if initiated in a particular way by some agent. One further important question concerns the nature of this probability judgment. How exactly should we interpret the probability in question? Is it some agent's credence that a project would be successfully realised if initiated in a particular way, say, the credence of the decision-maker in question? Or is it supposed to be an objective chance? There is room for some latitude here, depending on the precise function that the feasibility assessment is intended to play – perhaps it informs the advice an onlooker gives to an agent about how to act, or perhaps it is part of the deliberations of the agent him/herself. Whatever the precise details, it seems intuitive that feasibility is in some sense an objective matter. Perhaps it is best considered an objective epistemic probability – the credence that the decision-maker *should ideally* have, if appropriately informed, regarding whether the project in question would be successful if initiated in the specified way.[15]

2.2 Feasibility and self-interest

The characterisation of 'feasibility' above clearly does not *directly* concern self-interest, let alone Pareto improvements (which we return to in later sections).

[14] Admittedly, a group may count as a single agent if it is sufficiently unified in the relevant ways (as per List and Pettit 2011). But the conditions for agency are relatively difficult to satisfy, and one would be hard pressed to argue that the groups of nation states that concern us here count as single agents (even if each nation state separately counts roughly as a single agent).

[15] Of course, this raises further questions (which, again, come back to the precise function of the feasibility assessment) as to what it means for the decision-maker to be 'ideally/appropriately informed'. The broad account given here, however, is sufficient for the purposes of this discussion.

Feasibility is not a property of a value function *per se*; it is not a measure of the extent to which a theory or account of value is in line with self-interest. This point is worth elaborating. It is generally agreed that feasibility is not pertinent to the evaluation or ranking of states of affairs.[16] Thus if IP were understood as a theory of value (with Pareto improvements on the status quo ranked higher than states of affairs that are not Pareto improvements), then feasibility would not be a relevant consideration in assessing IP. More generally, when the question is one of evaluation—say, how we should evaluate different distributions of the costs of responding to climate change—feasibility is not relevant. Note that much discussion about the ethics of climate change concerns the aforementioned question of evaluation. It is only once we start talking about directives for action that feasibility becomes relevant, and as discussed, it concerns the likelihood that certain outcomes/events will come about, as opposed to the value of these outcomes.

Presumably when proponents say that 'IP is the most (or only) feasible approach to climate change', they are not making a purely evaluative statement but rather something in line with our working definition of feasibility above: That climate treaties designed to achieve a Pareto improvement on the status quo are likely to succeed in a way that climate treaties designed to involve sacrifices, are not. (IP proponents in fact tend to speak of feasibility as a binary matter, but as per the discussion above, a graded assessment is arguably more appropriate.) This is supposedly because the agents involved will only play their part in a treaty project if it furthers their self-interest. If true, this would amount to a more *indirect* relationship between feasibility and self-interest. It is an empirical claim about what reliably motivates the agents that climate projects depend upon.

Let us not forget that, throughout our discussion, the function of feasibility assessments is to aid a decision-maker. One thing to note is that, if we focussed just on whether it would be possible for a decision-maker to *initiate* a project in a specified way (whether this is a feasible act a la Southwood and co.), the association with self-interest would be tenuous. Recall why a simple act may not be a viable choice for an agent: If he/she would suffer some kind of irrationality or weakness of will that would ultimately prevent the performance of the act, *even if he/she preferred it and tried to choose it.* This does not let the agent off the hook easily when it comes to onerous acts involving self-sacrifice. In most cases, agents who act selfishly do not do so because it was not viable (or *feasible*, by the lights of Southwood and co.) for them to do otherwise. More likely a range of acts were viable, and yet the agent preferred the selfish one. To give an example, it is surely perfectly viable for me to set up a regular online donation to a charity. That is, if this were what

[16] Wiens (2015, 461) explicitly makes the distinction between directives for action and purely evaluative ethical claims, and notes that feasibility is pertinent to the former as opposed to the latter.

I preferred to do, I would succeed in doing it. If I do not set up the regular donation, it is rather because this is not in fact what I prefer to do, given my beliefs and values.

It may thus be perfectly viable for a decision-maker to initiate both more and less demanding climate treaties. But some of these treaty projects may be more likely to be fully realised than others, i.e., some of these treaty projects may be more feasible than others. This is the sense of feasibility that was argued above to be most apt for our discussion here. It is easy to see why this notion of feasibility is important for a decision-maker – it concerns how likely are the planned consequences of an initiating act – and moreover, why feasibility, thus understood, may be associated with self-interest. It all depends on the context and the relevant empirical facts. The likelihood that a complex project will be realised, if initiated in a specified way, depends on whether other agents the project depends on (perhaps even 'future selves' of the decision maker) will play their part.

The substantial question then is: Should a decision maker be more confident that a project or multi-stage plan she initiates would be realised to the extent that others are expected merely to act on their self-interest? Common wisdom suggests the answer is 'yes': Arguably, people more reliably act on self-interest than on other motivations. In that case, a project in which others are expected to act on self-interest as opposed to other motivations would be more likely to succeed, and would consequently be more feasible. But this picture of motivation may of course be overly simplistic. Moreover, it surely depends on the type of agent involved, whether an individual or a group entity such as a state. For now, let us simply note that the association between feasibility and self-interested motivations is an empirical matter, and one that is sensitive to context.

3 International paretianism and climate change

Putting aside, for now, the relationship between self-interest and feasibility, let us examine what the climate predicament looks like if cashed out in terms of the self-interest of the nation state actors. The first thing to note is that the self-interest of nation states – being large sprawling groups extending over time and space – is not a straightforward matter. As noted in Section 1, here we assume that the 'national interest' is an appropriate aggregate of the welfare (however this is cashed out) of present and future citizens (with considerable, if not equal, weight to future citizens). That is, we assume that there is some appropriate measure of the national interest of a state, as per the self-interest of an individual.[17]

[17] Note that Posner and Weisbach (2010) do not consistently treat the 'national interest' this way in their defense of the IP approach to climate change. What they mean by a nation's interest seems to vary, depending on whether they are discussing the feasibility merits or rather the moral merits of IP. When

The initial aim here, in Section 3.1, is to canvas plausible models of the climate-change predicament when posed in terms of the national interest of state actors. Whether or not state leaders are actually motivated to pursue the national interest, as roughly defined here, is a further issue that will be discussed in Section 3.2, along with other motivational concerns about climate treaties that aim for Pareto improvements. That is, the game models here do not necessarily track states' existing motivations (or rather, their apparent motivations/values given their choices). The models rather present one way of conceiving the climate-change predicament – where (the outcomes of) strategies are evaluated according to the national interest of the respective actors.

3.1 Prospects for IP climate treaties

This section considers what an IP climate deal might look like by appeal to key findings in public economics regarding the usage of 'common pool resources'. The key question is: What kind of game model plausibly fits the empirical facts and our rough definition of national interest? This is crucial for determining what sort of Pareto improvement is in the offing. Posner and Weisbach (2010, p. 6), for instance, suggest (in some places at least) that IP would support a global mitigation effort that is optimal (at least by utilitarian standards) and moreover substantial. We will see that this is not necessarily the case; indeed a number of factors affect the collective optimality and extent of mitigation under IP, as well as the nature of individual contributions.

The clearest reason for there being an unrealised opportunity for Pareto improvement on the status quo is a failure of collective action of some sort. Environmental resources are prone to collective action problems. Consider, for example, the over-exploitation of fisheries, deforestation, and of course climate change. The diagnosis: Many environmental resources such as those mentioned (abundant fisheries and forests, a stable climate, etc.) are *common pool resources* shared amongst agents who are primarily concerned with their own consumption of it. Technically speaking, common pool resources are effectively *non-excludable*, meaning that there are no barriers to anyone using the resource.[18] Common pool

---

it comes to feasibility, they seem to allow future welfare, for instance, to be weighed in the 'national interest' just as much as current citizens see fit (see, e.g., the discussion of the importance of respecting a state's sovereignty with respect to its own future in the discussion of the asteroid analogy on p.77). In many other places, however, they emphasize the ethical significance of the mitigation that would come from an IP treaty, suggesting a more ethically robust notion of 'national interest' as per our discussion in this chapter (see especially their comments on p.177 to this effect).

[18] Some environmental resources may also be *non-rivalrous* (at least up to some threshold), meaning that the resource in not depleted (up to the threshold) by additional users. Goods that are both *non-excludable* and *non-rivalrous* are referred to as *public goods*. (Strictly speaking, it is a matter of degree as

resources have a tendency (absent any institutional arrangement) to be over-exploited, because no one in particular has the responsibility/ability/motivation to provide or maintain the resource given that others cannot be excluded from using/spoiling it.

Climate stability is surely a classic example of a global common pool resource.[19] It is commonly assumed that the collective action problem takes the form of a Prisoners' Dilemma of international magnitude, where the 'cooperative solution' would be a clear Pareto improvement over the 'non-cooperative solution'. But this need not be the structure of the *game*, even assuming that the state actors pursue their 'national interest'. The Prisoners' Dilemma, considered particularly problematic when it comes to cooperation for mutual benefit, can be contrasted with 'mere' coordination problems. Moreover, even when it comes to Prisoners' Dilemmas, there is variation in the size of the Pareto improvement in question. In what follows, these different game scenarios are discussed in more detail: Section 3.1.1 briefly considers the coordination-game account of climate change, after which Section 3.1.2 turns to the Prisoners' Dilemma account and the variety of forms it may take.[20]

3.1.1 Climate change as a coordination game

Some suggest that, due to the presence of a dangerous threshold of emissions that would bring catastrophe to all, swamping other impacts, climate change may be best conceived as a *coordination game* (see Barrett 2011 for discussion).[21] Here the

to whether a good has these properties.) While the literature on environmental dilemmas often refers to *public goods*, it is generally only the property of non-excludability that is at issue, in which case, better simply to refer to *common pool resources*.

[19] Note that there are various ways to depict climate change as involving a common pool resource. Here we refer to the resource of 'climate stability', as per Ostrom (2009). Others refer to the 'carbon sink', or the Earth's ability to absorb greenhouse gas emissions, as the common pool resource, rather than a stable climate *per se*. The difference is not a substantive one.

[20] The models of this section follow the classic work of Barrett (1990) and Carraro and Siniscalco (1993) on international protection of the environment. Note that Gardiner (2001, Section VII) has a broadly similar discussion of the alternative ways to conceive of the intra-generational problem of environmental protection (primarily climate change); a key difference however, is that, in the first instance at least, Gardiner focuses on individual persons as the actors rather than nation states. It is worth noting also a further issue, which will not be explored here (in line with the literature on international environmental protection): How the structure of an existing game may be changed via new incentives. Ostrom (1990), for instance, has done very important work on the diversity of governance schemes that may resolve the over-exploitation of a common pool resource, by effectively changing the nature of the game. For the most part, however, the situation is such that, even if actors consent to a scheme for governing the resource, an external authority of some sort is required to ensure compliance with this scheme. It is typically assumed that such an authority is not available in the international setting.

[21] Typically the game is simplified such that there is negligible benefit to emissions abatement that is short/in excess of the designated threshold.

cooperative solution simply requires coordination: If all states were confident that the others were pursuing a common project for emissions abatement that just meets the threshold for avoiding catastrophe, then it would be in the interests of each to pursue this project as well. In other words, the presence of a dangerous threshold would, ironically, be good news for collective action, because there would exist arrangements for emissions abatement that are stable in the sense that no actor would have an incentive to defect. These sorts of joint strategies that resist unilateral defection are referred to in game theory as *Nash equilibria*; they are commonly regarded the 'solutions' to the game.

The presence of a dangerous climate-change threshold is unlikely to be as rosy as first appears, however, even from the perspective of collective action. To begin with, there is the problem of too many Nash equilibria. There would be countless projects or joint strategies involving just enough emissions abatement to avert catastrophe; some of these would involve heroic abatement efforts on the part of any given actor with little abatement from others, whereas some of them would involve very little abatement on the part of the actor in question and much greater efforts from others. It would be no trivial task for the actors to settle on one particular stable joint strategy – a tough bargain. As such, it might be risky for any given actor to pursue a particular joint strategy for averting catastrophe, given there is no guarantee that others would also opt for the same equilibrium strategy. The riskless Nash equilibrium would be for all to do little and suffer catastrophe without additionally engaging in costly and potentially futile abatement.[22]

An even more crucial issue, however, is that while climate change plausibly involves a danger threshold, there is uncertainty within the scientific community about its location. (The critical amount of 2 degrees C warming has gained traction in international debate, but apart from any other complications, there remains uncertainty about what *likelihood* for this temperature increase should be treated as the threshold, and there is furthermore uncertainty about the stock of atmospheric GHGs that corresponds to the chosen likelihood.) According to Barrett and Dannenberg (2012), this sort of uncertainty would effectively turn the prima facie coordination game into the more difficult Prisoners' Dilemma (PD) game. For this reason, our main focus here will be the PD account of climate change. Section 3.1.2 considers the variety of forms, when it comes to the magnitude of the effect (so to speak), that the climate PD could take.

[22] This is to suggest that the coordination problem has the form of the so-called 'stag Hunt' (see Skyrms 2004 for extensive discussion of this game). Others suggest the climate problem has the form of the roughly similar 'Battle of the Sexes' game (as discussed in Gardiner 2001).

3.1.2 Climate change as a Prisoners' Dilemma

It helps to begin with the simplest characterisation of how climate stability may give rise to a Prisoners' Dilemma. In our simple model, there are *N* relevant actors that affect the climate by (potentially) emitting greenhouse gases (GHGs).[23] Assume the *N* nations have equivalent circumstances and so the game is entirely symmetric (we return to this assumption later): Each actor can reduce or abate emissions, relative to the status quo or 'business as usual' (BAU), at a constant cost of *c* and a constant benefit of *b* per unit of abatement. Given that climate stability is a *common pool resource*, the costs of emissions abatement are private (borne just be the actor doing the abatement) but the benefits of this abatement accrue to all actors. So if each of *N* nations abates one unit of emissions, they each receive a benefit of *Nb*. The classic Prisoners' Dilemma results from the following pattern of costs and benefits:

$$Nb > c > b$$

Since, for each actor, the individual cost, *c*, of their own personal abatement effort is greater than the benefit, *b*, they would receive from this effort, there is no incentive to act. Whatever others do, it is better for the individual to do nothing. And yet, all individuals would benefit if they pulled together, and therein lies the dilemma: If all individuals were to similarly engage in emissions abatement, the benefit to each and every actor, *Nb*, surpasses their respective individual costs, *c*.

To use more formal language: In the PD game just described, the *non-cooperative* solution, where each actor chooses their level of emissions abatement based on their own interests, holding fixed what others do, amounts to zero total abatement and so zero total net benefit. This is because, for each actor, regardless of what others do, their own marginal contribution to the climate through abatement brings less private benefit *b* than the private cost *c*. This is the *Nash equilibrium* for the PD game—the solution that is stable in that no actor has an incentive to defect. The *fully cooperative* solution, on the other hand, is what brings maximum benefit to each actor; it is also optimal for the group in the aggregate welfare sense. Here it involves all actors pursuing maximum abatement *x*, which brings $(Nb - c)x$ net benefit to each actor. Note, however, that this Pareto improvement is not a (stable) Nash equilibrium as all actors have an incentive to defect and not abate at all.

A more realistic model has cost and benefit functions that are not constant but rather depend on levels of emissions abatement.[24] Assume $i = 1,..., N$ symmetrical

[23] Even though there are other ways to hasten/mitigate climate change (e.g. (de)forestation), GHG emissions play a key role and thus are our focus here.

[24] This is considered more realistic for a pollutant that does not accumulate in the atmosphere, unlike

actors as before. Let the total emissions abatement, Q, be the sum of all individual abatement efforts; so $Q = \sum_N q_i$, with q_i being the abatement of each actor. Assume that the marginal benefit of a unit of abatement for each actor, B_i, depends on total abatement as follows:

$$B_i(Q) = b(aQ - \frac{Q^2}{2})N \qquad \text{where } a \text{ and } b \text{ are positive parameters.}$$

This is a concave function with respect to global emissions abatement, which is to say that there are diminishing marginal returns for each additional unit of global abatement. The cost of abatement for each actor, C_i, is a convex function of its own abatement levels:

$$C_i(q_i) = cq_i^2/2$$

This amounts to increasing marginal costs of abatement: the more a nation individually abates, the more costly the next unit of abatement.

As per the previous model, the *non-cooperative* (Nash equilibrium) solution for non-constant cost/benefit functions is determined according to abatement levels that individual actors would choose, holding fixed the abatement of others. The solution, corresponding to total abatement Q_0, is such that the marginal abatement cost for each individual equals the marginal abatement benefit for that individual alone. The *fully cooperative* solution, on the other hand, corresponding to total abatement Q_C, is such that the marginal abatement cost for each individual is equal to the global marginal abatement benefit. That is, Q_C maximises the total benefit to the group, measured as the sum of benefits to individuals minus the sum of costs. The abatement and the net benefits for each actor under full cooperation are greater than the abatement and the net benefits for each actor under the non-cooperative solution; the difference, Q_C - Q_0, being the gain from full cooperation.

The size of this gain depends on the ratio c/b (the cost factor to the benefit factor) and the size of c (see Barrett 1994, pp. 880-1). When c is small and b is large, the cooperative gain is small because actors already benefit significantly from unilateral or non-cooperative abatement. When c is large and b is small, the cooperative gain is small because it makes little sense to abate, whether unilaterally or cooperatively.

GHGs. Barrett (1994) notes that Nordhaus (1990) employs a logarithmic cost function that depends on percentage emissions abatement, to model costs associated with GHG emissions abatement in the US. The precise forms of the cost and benefit functions, however, do not affect our (more general) discussion here.

The biggest gains from full cooperation, both with respect to total abatement and net benefits, are when *c* is comparable to *b* and both are large.

So we see that if the international climate predicament were best modelled as a Prisoners' Dilemma (PD), the size of the Pareto improvement (let alone whether or not it is achievable) is yet a further question. The model above well illustrates this point. For instance, in one of the PD scenarios mentioned just above—where the individual costs of emissions abatement are small and the benefits large—little is gained from a cooperative climate deal *per se*. Of course, such a model, if plausible, may nonetheless inspire change in revealing that the international community falls short even of the *non-cooperative* solution with respect to GHG emissions abatement; perhaps states need to start acting on their 'national interest', so to speak, rather than overcome a further collective action problem.[25] Some have indeed argued that it is not well enough appreciated that the benefits a single state receives from its own emissions abatement does to a large extent surpass the costs. Just because climate stability is a common-pool resource does not automatically mean that it is against a state's national interest to unilaterally do anything about it. There may well be indirect national benefits (or, in other words, lesser national costs than appreciated) associated with emissions abatement, such that the difference in abatement under the cooperative and the non-cooperative solutions is small. Ostrom (2009), for instance, emphasises local gains (or co-benefits) associated with emissions abatement, including savings in energy costs and health benefits from lower pollution (cf. Maslin and Austin 2012 on energy security and air quality).[26]

The other PD scenario where the gains of cooperation are slight, is when the cost factor, *c*, for emissions abatement is extremely large compared with the benefit factor, *b*. In this case, there is little reason, whether non-cooperatively or cooperatively, to engage in emissions abatement. This kind of model would only be plausible if, for each state, i) the BAU climate change path was relatively innocuous and/or ii) the welfare of future citizens was largely irrelevant to the 'national interest'. The first scenario is not very likely, given our best science. There is much uncertainty, possibly irresolvable, about how particular regions of the world will be affected by any particular rise in average global temperature, but nonetheless it is most plau-

[25] If this were the case, the supposed motivational power of the 'national interest' that is implicit in the defence of IP may be undermined, because one reason for states not currently pursuing their national interest and achieving the non-cooperative solution is that this is politically difficult. Section 3.2 will revisit this general issue of how easy it is for states to pursue their national interest.

[26] Others focus on the dynamics of the problem, and argue that the costs of emissions abatement decrease rather than increase, the greater the abatement effort to date (for various accounts of this phenomenon, see, for instance, Victor 2011, Heal and Kunreuther 2011, Tavoni 2013). The basic idea is that new partnerships and energy economies associated with emissions abatement would rapidly reach some critical mass whereby it is less costly to be in than out.

sible that the *expected* (in the mathematical sense) consequences are negative for all, compared to a future with a more stable climate.[27] The second reason is a non-starter in the context of our discussion here. It is true that, the more the future is discounted, when it comes to a state's 'national interest', the less benefit and the more cost associated with emissions abatement. But a heavily discounted future contravenes one of the premises of our discussion, namely that the 'national interest' of a state is some *appropriate* aggregate of the welfare of present and future citizens.

So it is important to note that the size of the cooperative gain associated with a PD can vary quite significantly. Thinking about these possibilities helps to sharpen one's views about the climate predicament that we face (based on 'national interest' as roughly defined here). There is reason to think that the status quo falls short even of a non-cooperative climate solution, that things would be different if states really acted in their national interest. On top of that, there is reason to think that the cost of GHG emissions abatement for individual actors is high, and the gains from cooperation also high. That is, it is not at all implausible that the climate predicament involves a Prisoners' Dilemma where the gain from cooperation would be substantial and thus tragic if not achieved. (It would arguably be more convenient if climate change gave rise to a 'mere' coordination game, but this seems wishful thinking.[28]) A further interesting question is whether the fully cooperative solution to the climate PD would effectively be what a global planner would recommend as the optimal mitigation response. This is a topic for further investigation, but the answer surely depends on whether the total welfare calculations of the global planner align with the total welfare calculations that underwrite the national interest of the respective states.[29] With the character of the Pareto improvement in mind, let us now consider whether a climate treaty that aims for this outcome is really achievable just on the back of ordinary 'self interested' motivations.

---

[27] See Frigg et al. (forthcoming) on the uncertainties surrounding predictions of *regional* climate change. For the multi-region models of climate economics investigated in Nordhaus and Boyer (2000), it is indeed the case that the expected consequences of BAU climate change for all regions of the world are negative, compared to selected mitigation options. Note too that the mitigation options are costly for the present generation in all regions of the world, compared to BAU. This makes clear that mitigating climate change is better than BAU (on these models at least) only because significant weight is given to the welfare of future generations (enough to counterbalance present costs).

[28] Posner and Weisbach (2010) often seem to assume, through their choice of analogies, that climate change is a coordination game. They do, however, also discuss the possibility of a Prisoners' Dilemma scenario (2010, 181).

[29] For instance, for a utilitarian global planner – as in the analyses of Stern (2007) and Nordhaus (1990) – the fully cooperative PD solution would only line up with that of the global planner if each state's national interest were also a utilitarian aggregate across all citizens in the present and future. This question is relevant to the 'efficiency' argument for IP that was briefly mentioned at the start of Section 2 (see, especially, footnote 2).

3.2 Does 'self-interest' suffice?

If the climate-change predicament were plausibly modelled as a game showing a large mutual gain for the state actors if only they could somehow rise to the cooperative task, then this would seem to be good news for International Paretianism. After all, IP is all about climate proposals that are *feasible* in virtue of appealing only to the cause of self-interest. And surely we can count on actors to merely pursue their self-interest (give or take some misconceptions they might have about what is really in their own self-interest). Even if we grant this simple rule of thumb about motivation, however, there is reason to doubt whether the Pareto improvements suggested by the Prisoners' Dilemma models canvassed above are achievable just on this basis. To begin with, we should be cautious in treating the 'national interest' of states akin to the self-interest of individuals. And secondly, even if states may be persuaded to pursue their national interest, as defined here, this alone does not suffice for overcoming a Prisoners' Dilemma. In what follows, these two issues are discussed in turn.

Many would observe that the analogy between the self-interest of an individual agent, both in explaining and guiding behaviour, and the 'self-interest' of a group agent, is far from tight. The group agents here are nation states. Do states naturally pursue their 'national interest', especially as understood here to involve due concern for the welfare of both present and future citizens? This seems a rather heroic assumption, and yet an important one for IP, because the appeal to 'self-interest' has force precisely because it is supposedly motivating in a straightforward way; something that an agent would quickly recognise as worth pursuing. The obvious worry is that state leaders rather have perverse incentives, when it comes to their domestic and international negotiations, having little to do with the pursuit of the 'national interest'. Moreover, even if democratic institutions provided some assurance that state leaders were responsive to the welfare of their citizens, it is a further stretch to think that this would somehow incorporate the welfare of future citizens.[30]

While there is a big question mark as to the motivating force of enlightened 'national interest' in international negotiations, let us nonetheless proceed on the basis that this is at least a persuasive reason for state action.[31] As noted already, it

[30] In response to this point, Posner and Weisbach (2010) might be tempted to loosen the notion of 'national interest' so that it is more in line with the existing apparent motivations of states. But, as noted already in footnote 13, this would conflict with their claims about IP solving the climate predicament; the climate solution they have in mind depends on an account of national interest that gives due weight to the welfare of both present and future citizens.

[31] Arguably, people are more naturally disposed to care for their own descendants as compared to contemporaries in distant parts of the world (cf. comment to this effect in Posner and Weisbach 2010, 142).

would no doubt be a significant international achievement if states were persuaded to act in their national interest, thus achieving the non-cooperative game solution. As far as securing the further mutual gains associated with the cooperative game solution – the promised Pareto improvement – there are, however, further motivational worries. The problem is that *projects* conforming to IP may require a lot more than the pursuit of national interest to succeed, once initiated. Here the lessons of game theory are important. The notorious problem with Prisoners' Dilemmas is that the fully cooperative solution is not a Nash equilibrium, and worse still, the greater the gains from cooperation, the greater the incentive for individual actors to free ride on the efforts of others, if the project were initiated. So even though the Pareto solution does further all actors' interests relative to the status quo, more than self-interested motivations are required to successfully implement this solution. Actors pursuing self-interest alone will try to free ride on the efforts of others.[32] To use the language of feasibility, if states really were purely concerned with their national interest and the game has a PD structure, the Pareto solution is not feasible.

In many contexts a PD game analysis serves as a plea to change the incentive structure of the game, perhaps via punishments for free riding.[33] But this is arguably not an option in the international setting, given the lack of an overarching authority that can facilitate such a change. [34] Economists have otherwise explored the possibilities that may arise from coalition agreements involving only some of the actors (here states) in the game. Solutions involving coalitions may be better than the Nash equilibrium and yet also dynamically stable in the sense that no actor included in the coalition has an incentive to leave, and no actor excluded from the coalition has an incentive to join. Hence these sorts of agreements are dubbed 'self-enforcing' within the public economics literature (see Carraro and Siniscalco 1993, Barrett 1994 & 2005).

Unfortunately, the cooperative possibilities associated with 'self enforcing' agreements are rather limited, as the aforementioned authors have shown. By way of elaboration: A stable or 'self-enforcing' coalition arrangement is one where the individual benefit to all members within the coalition is greater than it would be if they unilaterally defected (since the remaining coalition members would accor-

[32] Posner and Weisbach (2010) do acknowledge this point (see, esp., p.181), but they do not explore it in detail and nor do they give it appropriate emphasis.
[33] Recall the earlier reference to the work of Ostrom (1990).
[34] Although one might argue that a higher authority can at least be approximated in the international setting via multiple agreements, whereby punishments for non-cooperation in one domain (e.g. environmental protection) would take the form of sanctions in another domain (e.g. trade). See, for instance, Barrett (2008) for discussion. Nordhaus (2015) makes a similar point—that a groups of states or 'climate club' can effectively impose external trade-related penalties on states that do not cooperate with respect to emissions abatement.

dingly lower their own cooperative efforts, in this case emissions abatement), and moreover, the individual benefit to all those outside the coalition is greater than it would be if they were to become members (since the extra gains from further cooperation would not outweigh the individual costs).[35] Of course, it may be the case that all states would prefer to be outside the coalition, thus shouldering less of the total emissions abatement, than inside it. States within the coalition cooperate to achieve the greatest total benefit for the coalition, while the remaining states determine their abatement levels according to what is individually optimal, holding fixed the abatement of others. Once a stable coalition has formed, however, no state can benefit from unilateral change.

Across a wide variety of assumptions regarding the precise shape of the benefit and cost functions underlying a Prisoners'-Dilemma game, it turns out that little is achieved over and above the non-cooperative solution by stable coalition arrangements (as shown in Barrett 1994). The first game described above in Section 3.1.2, for instance, involving constant marginal cost and benefit functions, unsurprisingly does not support any self-enforcing coalition agreement. The second game, involving a concave benefit function and a convex cost function, supports only a limited advantage in coalition formation. For instance, when the number of stakeholders, *N,* equals 100, and the cost to benefit parameter ratio, *c/b*, equals one (so there is potentially much to gain from cooperation), the self-enforcing agreement involves only three states; these signatories abate considerably more than they would under the non-cooperative arrangement, while non-signatories abate just a little less. The large number of non-signatories means that the free-riding effect is substantial and so little is gained over and above the non-cooperative solution (see Barrett 1994, p. 882).[36] In fact, to the extent that a coalition in this class of cases has a large number of signatories, there is little to be gained from cooperation (*c/b* is either very large or very small). Similar results apply to other functional forms: generally speaking, where there is much cooperative gain at stake, only coalitions with 2 or 3 signatories are stable, and thus not much cooperative gain can actually be realised in a 'self-enforcing' manner.[37]

[35] The notion of a stable coalition is attributed to earlier work on cartels by d'Aspremont et al. (1983).
[36] Note that these results are obtained via simulation. For other cost and benefit functional forms, analytic results can be derived.
[37] One might wonder whether the situation changes if we allow asymmetry, i.e., when cost and benefit functions for the *N* states are not identical. There has been some investigation of this case in the literature. While Barrett (1997) derives similarly pessimistic results for heterogeneous as for homogeneous parties to an agreement, McGinty (2007), on the other hand, finds that, assuming transfers are permitted within the coalition, heterogeneity can significantly increase the percentage gain from full cooperation, even when this gain is substantial. It seems that further investigation is required to establish the robustness of this result.

4 The feasibility of IP climate deals

The assumption that states pursue their national interest may yield a PD climate game with a sizeable Pareto improvement, but we see that 'self-interest' alone may not get us very far in realising that Pareto improvement. IP proponents would thus be better served by a more nuanced story regarding the association between feasibility and self-interest. In fact, proponents do try to sell IP on moral 'other-regarding' grounds as well as 'self-interest' grounds. Throughout their book, Posner and Weisbach, for instance, appeal to the happy confluence of gains in national interest and moral gains under an IP climate treaty. This suggests a hidden commitment to the moral standing of a project being relevant to its motivational appeal and thus feasibility, even if the moral gains are presented as if they were a mere by-product of the pursuit of self-interest. Recall too that the pursuit of 'self-interest' in the context of IP is already morally demanding. It is hardly the norm for state leaders to act in the interests of present and future citizens, suitable balanced. So IP proponents appeal in various ways to moral reasons for climate action, without necessarily advertising this fact.

Posner and Weisbach make an even larger concession regarding the need to appeal to moral reasons for climate action. They state that motivations beyond 'self-interest' are required to secure Pareto improvements where states have an incentive to free-ride (p. 169):

> A key for a climate treaty is what we have called International Paretianism—nations must believe that they are better off with a treaty than without. But the obligation to achieve a broad, deep, and enforceable treaty imposes a serious ethical duty on rich and poor nations alike—the obligation to cooperate. In our view, it is unethical for a nation to refuse to join a climate treaty in order to free-ride off of others.

While it has an air of authority, this statement of ethical duties has no clear justify-cation.[38] The question, when it comes to the feasibility merits of IP, is whether a climate treaty relying on this extra commitment to secure a Pareto improvement would be *more or less motivationally compelling*, for all actors concerned, as a treaty relying just on the pursuit of national interest.

Let us briefly consider this question. Refraining from free-riding involves a sacrifice in national interest. For this to be motivationally compelling, it would need

[38] As acknowledged by Posner and Weisbach (2010, 180), although they do not allow that this lack of justification severely undermines the feasibility case for IP.

to be offset by some other, say moral, gain. The question is then: Does an IP climate deal have the required moral merit? Does it make an obvious step forward in moral terms? On certain ways of calculating national interest and the global good, the IP climate solution may well accord with the optimal amount of global climate change mitigation. But note that this is consistent with the present generation of poor countries paying deeply for this mitigation so as to ensure climate stability for their descendants. (The alternative, of course, would be for other nations to pay more for the same mitigation effort.) So we see that, in solving one global problem – climate stability – another is exacerbated, namely, inequalities (not to mention the failure to meet historical responsibilities). Does it make sense for nations on the losing side of this equation to accept an imperative to participate in an IP climate deal rather than hold out for an alternative mitigation proposal that better promotes equality? It is far from clear that IP has the upper hand here, in terms of being motivationally compelling.[39]

More precisely, the problem for IP proponents is that, once we allow for other moral motivations to weigh against the pursuit of national interest, the feasibility case for IP is much shakier. This is exacerbated when the gains from emissions abatement are not symmetric and when states have more complex moral concerns (say, they care about equality and/or historical responsibility). In this case, states may have very different views about the moral merit of an IP climate treaty, which may well affect the overall appeal of the proposal for them. For rich, high-emitting states, an IP treaty is good for national interest and it may look pretty good morally-speaking as well, when one compares to the business-as-usual baseline; thus rich states may have both national interest and moral reasons to commit to an IP treaty. But poor states may rather see an IP treaty as morally deficient, compared to other possible treaties that would better promote equality and compensation for past wrongdoing. In other words, they may see little reason to treat business-as-usual as the *moral baseline,* and thus have little moral reason to commit to an IP treaty. In that case, there would be little reason for them to refrain from free riding or defecting from such a treaty.

Let us then conclude on a cautionary note. While concern for the feasibility of climate treaties is important and timely, there is no simplistic way to make such assessments. Feasibility is simply the likelihood, from the deliberator's perspective, that a proposed project or multi-stage plan will be successfully implemented, once initiated in a specific way in a given context. When it comes to climate treaties, the

[39] See Brennan and Sayre-McCord (2016) for discussion of how moral facts can affect the feasibility of a political proposal. They make the point that the position of those advocating a mutually beneficial or 'no sacrifice' climate treaty on feasibility grounds is inconsistent, at least when coupled with the claim that the broader problem of global injustice can always be addressed later (as per remarks of Posner and Weisbach, e.g. p.92).

deliberator (whether a member state or an outside official) must assess whether other actors, namely states, will be motivated to follow through on their part in the project, if it were initiated as well as possible by the deliberator. The more reliable these motivations, the more likely the project would succeed, i.e., the greater its feasibility. We must conclude then that it is an open empirical question as to whether states' motivations would likely align with an IP climate treaty; indeed it is plausible that states' motivations would not be so aligned, given that self-interest alone may be self-defeating, and states may well perceive the moral situation rather differently.[40]

Bibliography

D'Aspremont, Claude, Alexis Jacquemin, Jean Jaskold Gabszeweiz and John A. Weymark. 1983. "On the Stability of Collusive Price Leadership." *Canadian Journal of Economics* 16: 17–25.

Baer, Paul. 2013. "Who Should Pay for Climate Change? "Not Me"." *Chicago Journal of International Law* 13(2): 507–525.

Barrett, Scott. 1990. "The Problem of Global Environmental Protection." *Oxford Review of Economic Policy* 6(1): 68–79.

Barrett, Scott. 1994. "Self-Enforcing International Environmental Agreements." *Oxford Economic Papers* 46: 878–894.

Barrett, Scott. 1997. "Heterogeneous International Environmental Agreements." In *International Environmental Negotiations: Strategic Policy Issues*, edited by C. Carraro. Cheltenham: Edward Elgar.

Barrett, Scott. 2005. "The Theory of International Environmental Agreements." In *Handbook of Environmental Economics Volume 3*, edited by Karl-Goran Mäler and Jeffrey R. Vincent, 1457–1516. Elsevier B.V.

Barrett, Scott. 2008. "Climate treaties and the imperative of enforcement." *Oxford Review of Economic Policy* 24(2): 239–258.

Barrett, Scott and Astrid Dannenberg. 2012. "Climate negotiations under scientific uncertainty." *Proceedings of the National Academy of Sciences* 109(43): 17372–17376.

Brennan, Geoffrey and Geoffrey Sayre-McCord. 2016. "Do Normative Facts Matter... to what is feasible?" *Social Philosophy and Policy* 33(1–2): 434–456.

[40] Many thanks to Christian Barry, Richard Bradley, Fergus Green, Jeremy Moss, Nicholas Southwood, Kai Spiekermann, and all three editors of the *Philosophy and Climate Change* collection for very helpful comments on earlier drafts of this chapter.

Brennan, Geoffrey and Nicholas Southwood. 2007. "Feasibility in Action and Attitude." In *Hommage a Wlodek: Philosophical Papers Dedicated to Wlodek Rabinowicz*, edited by T. Ronnow-Rasmussen, B. Petersson, J. Jonefsson, and D. Egonsson. url: www.fil.lu.se/hommageawlodek, only published online.

Broome, John. 2010. "The most important thing about climate change." In *Public Policy: Why Ethics Matters*, edited by Jonathan Boston, Andrew Bradstock and David Eng, 101–116. Canberra: ANU E Press.

Carraro, Carlo and Domenico Siniscalco. 1993. "Strategies for the international protection of the environment." *Journal of Public Economics* 52: 309–328.

Estlund, David. 2011. "Human Nature and the Limits (if Any) of Political Philosophy." *Philosophy and Public Affairs* 39: 207–37.

Frigg, Roman, Leonard A. Smith, and David A. Stainforth. Forthcoming. "The Myopia of Imperfect Climate Models: The Case of UKCP09." *Philosophy of Science*.

Frisch, Mathias. 2012. "Climate Change Justice." *Philosophy and Public Affairs*, 40(3): 225–253.

Gilabert, Pablo and Holly Lawford-Smith. 2012. "Political Feasibility: A Conceptual Exploration." *Political Studies* 60: 809–825.

Heal, Geoffrey and Howard Kunreuther. 2011. "Tipping Climate Negotiations." Working Paper 16954, National Bureau of Economic Research.

Jamieson, Dale. 2013. "Climate Change, Consequentialism, and the Road Ahead." Chicago Journal of International Law 13(2) 439–468.

Lawford-Smith, Holly. 2013. "Understanding political feasibility." Journal of Political Philosophy 21: 243–259.

List, Christian and Philip Pettit. 2011. Group Agency: The Possibility, Design, and Status of Corporate Agents. Oxford: OUP.

Maslin, Mark and Patrick Austin. 2012. "Climate models at their limit?" Nature 486: 183–184.

McGinty, Matthew. 2007. "International Environmental Agreements Among Asymmetric Nations." Oxford Economic Papers 59(1): 45–62.

Nordhaus, William D. 1990. "To Slow or Not to Slow: The Economics of the Greenhouse Effect", mimeo, Department of Economics, Yale University.

Nordhaus, William D. 2015. "Climate Clubs: Overcoming Free-Riding in International Climate Policy." American Economic Review 105(4): 1339–70.

Nordhaus, William D. and Joseph Boyer. 2000. Warming the World: Economic Models of Global Warming. Cambridge, MA: The MIT Press.

Ostrom, Elinor. 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge: Cambridge University Press.

Ostrom, Elinor. 2009. “A Polycentric Approach for Coping with Climate Change.” Working Paper 5095, World Bank.

Posner, Eric A. and David Weisbach. 2010. Climate Change Justice. Princeton, NJ: Princeton University Press.

Schokkaert, Erik and Johan Eyckmans. 1998. “Greenhouse negotiations and the mirage of partial justice” In Global Environmental Economics, edited by Mohammed H. Dore and Timothy D. Mount, 193–217. Oxford: Basil Blackwell.

Shue, Henry. 2013. “Climate Hope: Implementing the Exit Strategy.” Chicago Journal of International Law 13(2): 381–402.

Skyrms, Brian. 2004. The Stag Hunt and the Evolution of Social Structure. Cambridge: Cambridge University Press.

Southwood, Nicholas. 2016. “Does “Ought” Imply “Feasible”?” Philosophy and Public Affairs 44(1): 7–45.

Stern, Nicholas. 2007. The Economics of Climate Change: The Stern Review. Cambridge: Cambridge University Press.

Tavoni, Allesandro. 2013. “Building up cooperation.” Nature Climate Change 3: 782–783.

Victor, David G. 2011. Global Warming Gridlock: Creating More Effective Strategies for Protecting the Planet Cambridge: Cambridge University Press.

Wiens, David. 2015. “Political Ideals and the Feasibility Frontier.” Economics and Philosophy 31: 447–477.

Göran Duus-Otterström[1]

Sovereign States in the Greenhouse: Does Jurisdiction Speak against Consumption-Based Emissions Accounting?

The paper investigates the significance of jurisdiction for the choice of accounting method of greenhouse gases. Making use of the distinction between retrospective and prospective responsibility, it assesses three different arguments from jurisdiction against consumption-based emissions accounting. It argues that one of these arguments, the effectiveness argument, provides a strong potential reason against consumption-based emissions accounting. To the extent jurisdictional control is needed to reduce some emissions, and production-based accounting incentivizes states to reduce these emissions, there is a reason of environmental effectiveness for sticking with production-based accounting.

[1] Aarhus & Institute for Futures Studies, gdo@ps.au.dk. Financial support by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences) is gratefully acknowledged.

1. Introduction

A key question of international climate policy is how emissions should be counted. The current method, known as *production-based accounting*, counts emissions at their geographical point of origin.[2] The United Nations Framework Convention on Climate Change (UNFCCC) instructs countries to keep inventories over 'greenhouse gas emissions and removals taking place within national territory and offshore areas over which the country has jurisdiction' (IPCC 2006, p. 1.4). An alternative method is *consumption-based accounting*. Under consumption-based accounting, a country is ascribed emissions embodied in the goods and services it consumes, but not emissions embodied in the goods and services it exports for consumption elsewhere. Thus, rather than counting the emissions that went into making, say, a smart phone in the countries that manufactured and assembled it, this method counts the emissions in the country where the final consumer resides (Peters 2008; Peters and Hertwich 2008; Davis and Caldeira 2010; Barrett *et al.* 2013; Lininger 2015).

Proponents argue that consumption-based (CB) accounting would provide for a more just and environmentally effective allocation of climate burdens. Counting emissions on the consumption side would shift emissions from developing to developed countries, and since developed countries are not only rich but also tend to have more ambitious climate policies, making them shoulder a greater share of the overall climate burden would be both appropriate and better for the climate.[3] This line of argument rests on two key assumptions. First, countries' levels of emission are relevant for their climate burdens. Second, countries would not lower their climate policy ambitions as a reaction to being ascribed more emissions. While the first of these assumptions is very plausible—no matter what we might think of it in principle, the way emissions are ascribed to countries does matter for their climate burdens in practice—the second is more doubtful. Still, the case for CB accounting is strong enough to have generated increasing interest in this option, and scholars have consequently debated its complexity, feasibility, justice, and effectiveness.[4]

---

[2] Some prefer the term 'territory-based accounting' for UNFCCC's system of accounting and use 'production-based accounting' in a slightly different way (Barrett *et al.* 2013). See also the distinction between production, consumption, extraction, and income-based approaches in Steininger *et al.* (2016).

[3] This is typically cast in terms of CB accounting preventing *carbon leakage* (the process whereby emissions relocate to countries where they are subject to less stringent regulation). A popular distinction is between *strong* (climate-policy induced) and *weak* (consumption induced) carbon leakage (Afionis *et. al.* 2017, pp. 3-4). See also the distinction between the relocation, energy market, income, and technological spillover channels of carbon leakage in Steininger et al. (2016).

[4] For summaries of the arguments for and against CB accounting, see Steininger *et al.* (2014) and Afionis *et al.* (2017). Roser and Tomlinson (2014) offer an overview of the arguments for and against the closely related idea of border carbon adjustments. The feasibility of CB accounting is discussed by Grasso and

One striking feature of the choice of accounting method is that international climate policy revolves around sovereign states or governments.[5] Environmental policies have traditionally been 'production-based' precisely because this allows states or governments to address problems in the territory over which they have jurisdiction. Yet CB accounting ensures that 'a part of the emissions occur outside of the regulated country where there is no jurisdiction to regulate emissions' (Peters et al. 2016, p. 51). For example, CB accounting would attribute some emissions produced in China to other countries even though the latter lack jurisdiction over Chinese territory. It is natural to suspect that this is a deep flaw of CB accounting. Despite this, the significance of jurisdiction has not been subjected to systematic scrutiny in the emerging normative debate on this method.

To rectify this shortcoming, in this paper I investigate what I call *the argument from jurisdiction against CB accounting*. The general idea behind this argument is simple: when attributing emissions, we should attribute them to the entity that has jurisdiction over the territory from which the emissions emanate, because doing so ensures that responsibility is assigned to the entity with the legal power to control emissions. In one version of the argument, this is important because it would be *unfair* to assign emissions otherwise (it would hold states responsible for emissions they did not cause or cannot control). In another version, it is important because it would be more *effective* (it would give the state with the best ability to lower emissions an incentive to do so). Using the tools of normative ethics and classical logic, my aim in this paper is to assess these arguments. I will suggest that while the arguments seem persuasive at first glance, a closer analysis reveals that they are subject to significant doubt. The fairness-based arguments are largely unsuccessful. The effectiveness-based argument provides a potentially strong ground for counting emissions at the point of production, but its soundness depends on empirical premises the truth of which remain to be established.

Three preliminary remarks about the paper: First, although emissions accounting is a fascinating and important topic in its own right, the argument from jurisdiction of course has wider implications than just emissions accounting. The argument is relevant for any context in which we are unsure about how to allocate responsibility between producers and consumers.[6] Second, while this paper given its focus on jurisdiction and legal control treats states or governments as the relevant bearers of climatic responsibility, nothing prevents applying some of the

Roberts (2014) and Grasso (2017). For good discussion of the additional complexity introduced by CB accounting, see Peters (2008).

[5] I use 'state' and 'government' interchangeably to describe the national political authority of a country, but for the sake of simplicity, I shall mostly speak of 'states.'

[6] For consumer responsibility and ethical consumption generally, see, e.g., Hussain (2012) and Lawford-Smith (2015). For a well-known analysis of responsibility under global supply chains, see Young (2006).

logic of the arguments in a more individualistic setting. One of the points I make below is that whenever we face a question of allocating costs between producers and consumers, we may think about it through the lens of both retrospective and prospective responsibility. That point remains relevant when we conceive of consumers and producers as individuals as opposed to states or governments. Third, I treat CB accounting and production-based (PB) accounting as rival and exhaustive options, meaning that if an argument speaks against one method, it simultaneously speaks in favor of the other. I do not consider other methods of accounting, nor do I consider potential ways of combining CB accounting and PB accounting.[7]

2. Retrospective and prospective responsibility for emissions

Before we get to the argument from jurisdiction, in this section I establish the distinction between retrospective and prospective responsibility. The debate on CB accounting is to large extent about responsibility for emissions, yet it does not sufficiently distinguish backward-looking and forward-looking senses of responsibility. Parsing the concept of responsibility more carefully will allow us to sort the argument from jurisdiction into its different versions.

It is widely recognized by philosophers that 'responsibility' does not refer to a unified concept. Vincent (2011) offers a taxonomy over no less than six responsibility concepts—responsibility as virtue, role responsibility, outcome responsibility, causal responsibility, moral responsibility, and responsibility as liability—each of which contain subtle distinctions in turn. For our purposes here it is enough to use a somewhat crude distinction between being responsible for something having occurred (*retrospective responsibility*) and being responsible for ensuring that something happens (or does not happen) in the future (*prospective responsibility*). The generic phrase 'Agent *A* is responsible for state of affairs *X*' is neutral between these two senses of responsibility. When we are ascribing retrospective responsibility, we are essentially trying to explain why an outcome occurred, often with an eye to assigning praise or blame. Prospective responsibility, by contrast, is about assigning tasks or duties. Here we are trying to identify the party or parties that have the task of ensuring that something happens (or does not happen) in the future.[8]

[7] For shared responsibility/costs models, see Lenzen et al. (2007).

[8] Retrospective responsibility is always about causal responsibility and in addition typically takes on the texture of *moral* responsibility in the sense that the retrospectively responsible actor is praise- or blameworthy for an outcome. Prospective responsibility is similar to what David Miller calls 'remedial responsibility', although it is not necessarily a special responsibility (Miller 2007). For prospective responsibility, see Goodin (1985) and van de Poel (2011).

It should be clear that, while the two senses of responsibility are distinct, they often interrelate. For example, when someone has littered your backyard, it is natural to argue that the person who caused the mess (retrospective responsibility) should clean it up (prospective responsibility). The influential Polluter Pays Principle connects the two types of responsibility precisely in this way.[9] Yet the two types of responsibility are logically distinct, and they can come apart in practice. For example, it is common to argue that the most capable agents should sometimes assume the responsibility for addressing a problem even though they did not bring it about. In such situations, retrospective responsibility is neither a necessary nor a sufficient condition for prospective responsibility.

The relevance of this distinction for the choice between CB accounting and PB accounting is that an accounting method could be seen as a way to ascribe either retrospective or prospective responsibility for emissions. Read as a way to ascribe retrospective responsibility, an accounting method is a way of saying who brought about emissions. Read as a way to ascribe prospective responsibility, however, it is rather a way of saying where the task of managing and reducing these emissions lies—to put it colloquially, whose business these emissions now are.

That the accounting method could be seen as serving different goals is important because it means that there are different ways of *evaluating* whichever distribution of emissions that an accounting method produces. More specifically, we can worry about this distribution for backward-looking and forward-looking reasons. Suppose one accounting method attributes x tons of CO_2 to a country C whereas another attributes y tons of CO_2 to C, and y is greater than x. Which attribution should we prefer? One natural question is whether one method is more accurate than another method from the perspective of who *caused* the emissions. For example, if an accounting method attributes y rather than x tons of CO_2 to C, then we may worry that this exceeds the emissions that C actually brought about. This is the same as using retrospective responsibility as the evaluative lens. But another question is whether one method is better than the other when understood as a way to allocate the task of reducing emissions, which is the aspect highlighted by prospective responsibility. For example, we may wonder whether a method is more likely to lead to emission reductions. The lesson is that we must always assess the way emissions are attributed to countries in light of some normative standard.

This is not to deny that we generally think of the accounting method in a particular way. It seems to me that we typically regard the emissions accounting

[9] However, different versions of the Polluter Pays Principle explain this in different ways. Read as a 'fault-based' principle, polluters should pay because this is just (Shue 1993). Read as a forward-looking principle, polluters should pay to ensure that would-be polluters have an incentive not to pollute (de Sadeleer 2002). For a philosophical discussion of the principle as applied to climate change, see Caney (2005) and Roser and Seidel (2017, pp. 118–129).

exercise as ascribing retrospective responsibility *in order to* assign prospective responsibility. Our assumption then is (i) that attributing emissions to countries is the same thing as giving them the task of managing or reducing emissions and (ii) the task of managing or reducing emissions should befall those who are retrospectively responsible for them. But the backward-looking and forward-looking concerns are distinct and could be valued independently. Those who see it as unimportant that emissions are attributed in accordance with retrospective responsibility, for example, will not worry about that a method might attribute emissions to countries that did not cause those emissions. Since an accounting method might be worse in one dimension but not the other, it is important to be clear about which dimension we invoke in assessing it.

3. The argument from jurisdiction

Let us now turn to the argument from jurisdiction against CB accounting. The argument relies on the claim that states have *effective jurisdiction* over their territories. 'Jurisdiction' should here be understood as the exclusive formal right 'to make and enforce laws throughout the territory' (Miller 2012, p. 253).[10] 'Effective' jurisdiction, meanwhile, refers to a capacity to *exercise* this right. Thus, the claim that states have effective jurisdiction over their territories amounts to the claim that states have the right and the ability to decide which laws and regulations to enact. It is this fact, so the argument from jurisdiction holds, that ensures that CB accounting would be a poor choice for international climate policy.

I shall not dispute the claim that states have effective jurisdiction over their territories. I recognize that there are cases where the claim is false, but the argument from jurisdiction only relies on that it is true for *most* states, and this seems plausible enough. Hence, it is no objection against the argument that there are individual cases where a state lacks effective jurisdiction because, for example, it has been annexed by another state or is ridden by civil war. It is also important to note that the argument does not assume that effective jurisdiction renders states omnipotent over what transpires in their territories. It only stresses that effective jurisdiction ensures that states alone are able to decide which laws and regulations to put in place. This does give states a special kind of control over what transpires on their

[10] By saying that it is 'formal' right, I want to signal that this is not a normative claim: I am not suggesting that all states that have jurisdiction over their territory *should* have it. Moreover, the right is 'exclusive' in the sense that the right to make and enforce law is only held by the state. The latter may seem too strong because, as one commentator notes, 'the Hobbesian model of a single sovereign with a unified, complete jurisdiction has never existed' (Valverde 2014, p. 387). But the argument from jurisdiction does not assume complete jurisdiction, so I will here gloss over the extent to which, if at all, the national state or government shares the right to make and enforce law with other political authorities.

territory, but it does not give them anything like complete control, nor does it rule out that other agents, like private business or foreign powers, may in turn influence their legislative choices.

Effective jurisdiction provides one reason to ascribe retrospective responsibility to the state for what transpires on its territory. Although individuals, firms and other organizations decide what to do *within* the framework set out by the legal order, so this thought goes, the state controls their actions in the sense that it controls the legal framework. For example, in the case of climate change, we could argue that the emissions emanating from a territory are ultimately the state's fault since the state establishes the framework within which these emissions were permissible (Caney 2005, p. 755). This is complicated once we factor in that no state is able to secure full compliance with laws and regulations. The point here, however, is that jurisdiction can and does ground claims for holding states' retrospectively responsible for what transpires on their territory. If a state is unable to enforce its laws and regulations, then this just means that it has little retrospective responsibility for what individuals, firms and other organizations do on its territory.

Jurisdiction can also play a key role for the ascription of prospective responsibility. Suppose it is difficult to lower the emissions emanating from a territory without having effective jurisdiction over that territory. The fact that a state has effective jurisdiction then makes it a natural site for prospective responsibility: we should ascribe the task of reducing emissions to the state because the state is the entity that is most able to affect the amount of emissions flowing from a territory. Thus, the same considerations that make the ascription of retrospective responsibility appropriate may underpin a case for ascribing prospective responsibility. Legal power and prospective responsibility should go hand in hand.[11]

The argument from jurisdiction flows from these observations. It holds that CB accounting is a bad idea because, in counting emissions at the point of final consumption, it attributes some emissions to states that lack the legal power to control them. Precisely why this is a bad idea comes in different versions, each of which could stand on its own but could easily be combined.

4.1 The fairness argument from retrospective responsibility

One argument is that effective jurisdiction is a precondition for, or strongly contributes to, retrospective responsibility for emissions. Suppose the United Kingdom

[11] Again, this is not to suggest that the amount of emissions flowing from the territory is *fully* under the control of the state. Individuals and other agents could choose to reduce emissions more than legally required or emit more than legally permitted. The point is simply that the state is in the best position to affect the amount of emissions in virtue of setting up proscriptions and prescriptions for other agents.

(UK) is attributed with the emissions that go into the goods it imports from China. According to the fairness argument from retrospective responsibility, the UK should not be held accountable for emissions associated with its consumption of Chinese goods since, given its lack of jurisdiction over those emissions, it had no control over how those goods were produced.

The argument can be given a positive or a negative form. The positive formulation draws on the claim that it is appropriate that those who cause a bad outcome are held accountable. Thus, what is problematic about CB accounting is that it inappropriately exonerates states from emissions associated with their exports. We can state this argument as follows:

> P1. It is appropriate to hold agents accountable for bad outcomes they cause.
>
> P2. CB accounting fails to hold agents accountable for some bad outcomes they cause.
>
> Therefore, CB accounting is inappropriate.

The negative formulation states, more modestly, that what is problematic about CB accounting is that it holds states accountable for outcomes it did not cause. While it is not important that those at fault are held accountable, it is important that no one is held accountable without being at fault. We can state this argument as follows:

> P1. It is unfair to hold agents accountable for outcomes they did not cause.
>
> P2. CB accounting holds agents accountable for outcomes they did not cause.
>
> Therefore, consumption-based accounting is unfair.

These arguments capture a natural objection to CB accounting, as evidenced by the emerging debate on whether this method of accounting conflicts with the Polluter Pays Principle (Duus-Otterström & Hjorthen 2018). Skeptics argue that CB accountting is incompatible with this principle since counting emissions at the point of final consumption exonerates the polluters, that is, those who actually *release* greenhouse gases into the atmosphere. As one commentator puts it, 'producers are principally, logically and obviously responsible for emissions from production' (Liu 2015, p. 5).

The problem is that premise P2 is doubtful in both arguments. The gist of these premises is that those who produce the emissions are the 'real' polluters, but it is not clear that this stands up to critical scrutiny. As Roser and Tomlinson (2014) note, the central concern for the Polluter Pays Principle is causation—polluters are those who cause emissions to occur.[12] Yet no theory of causation suggests that only producers are polluters thus understood. Both producers and consumers contribute to the causal sequences that give rise to emissions, and for both of them is true that, had they not behaved the way they did, the emissions would not have occurred. Thus, producers and consumers *jointly* cause emissions. The Polluter Pays Principle therefore neither favors nor condemns PB accounting. Instead, it holds that the emissions—and, more importantly, the costs associated with them—should be split between producers and consumers, perhaps in proportion to their degree of causal contribution.

We may quarrel that Roser and Tomlinson overlook the normative and conventional dimensions of posing a question like 'who is causing pollution?' What counts as a cause of an outcome cannot simply be read off the empirical facts but depends, as J.L. Mackie showed years ago, on our perspective or explanatory interest (Mackie 1965).[13] In explaining why pollution occurs, many different causes may be highlighted, ranging from high-level phenomena like human nature or capitalism, to more mundane factors such as the lack of sufficient alternatives to fossil fuels, and while these explanations are not mutually exclusive, different people will emphasize different ones depending on which factors seem salient to them. Hence, even though consumer demand is a necessary condition for many emissions, it does not follow without further argument that it counts as a *cause* of emissions. Perhaps we regard consumer demand as a mere background factor, much like the presence of oxygen in Mackie's famous example of a house burning down due to a faulty circuit. If so, we would hardly think of consumers as 'causing emissions.'

There is little doubt, however, that both consumers and producers are salient in explaining why pollution occurs. Indeed, it is precisely because consumers and producers seem to contribute to climate change that the question of apportioning retrospective responsibility between them has become the subject of discussion.

[12] To be precise, Roser and Tomlinson argue that the Polluter Pays Principle ultimately cares about *agent responsibility* (Vallentyne 2008), which is similar to the perhaps better-known concept of *outcome responsibility* (Miller 2007, pp. 86–97). However, they think that the question of states' responsibility for emissions essentially boils down to causal contribution since producers and consumers are typically agent responsible for their involvement in production-consumption sequences (Roser and Tomlinson 2014, p. 237).

[13] Mackie expressed this in terms of causal statements presupposing a 'causal field'. See also Miller (2007, p. 87).

Taking Mackie's more nuanced perspective on causal attribution therefore does not change the conclusion that producers and consumers are jointly causing pollution.[14]

The upshot is that the fairness argument from retrospective responsibility fails. Even though producing goods lies closer to what we intuitively think of as 'causing emissions', on reflection it is not obvious that consumers are any less of a cause of emissions than producers. Hence, we cannot say that attributing retrospective responsibility to them is unfair.

4.2 The rairness argument from prospective responsibility

Let us return to the example of attributing some of the emissions emanating from China to the UK. We just saw that it is not obviously unfair or inappropriate to hold the UK retrospectively responsible for these emissions since the UK can be seen as a cause of them. Another perspective on the issue, however, does not concern the fairness of blaming the UK for emissions it purportedly did not cause, but rather whether it is fair to assign the UK the prospective responsibility to reduce these emissions *moving forward*. Note that this concern is independent of whether the UK is retrospectively responsible for the emissions or not.

The fairness argument from prospective responsibility offers a different and much less discussed way to attack CB accounting. We can state this argument as follows:

> P1. It is unfair to ascribe prospective responsibility to an agent when the agent cannot discharge that responsibility.
>
> P2. Consumption-based accounting ascribes prospective responsibility to agents that cannot discharge that responsibility.
>
> Therefore, consumption-based accounting is unfair.

What can we say about this argument? Premise P1 is very plausible, falling under the well-known rule that 'ought implies can.' If CB accounting means that the UK is supposed to lower emissions over which it has no control, then CB accounting would clearly be unfair to the UK.[15]

[14] It does not follow from this that they are *equally* causally responsible. It is possible to argue that *A* and *B* are jointly responsible for *X* yet argue that *A* is *more* responsible than *B*. Indeed, this is normal in many cases of pollution, such as when two cities together pollute a lake, but one city releases more pollutants than the other. But this qualification does not affect my analysis as any degree of retrospective responsibility on the side of consumers is enough to defeat the fairness argument from retrospective responsibility.

[15] China might nevertheless act in a way that conforms to what the UK should have brought about. In

Premise P2, however, is more doubtful. First, it is far from self-evident that the UK does lack control over Chinese emissions. What the premise gets right is that the UK is unable to make *legal* decisions in the Chinese territory. Factories in China obey Chinese law, not British law. However, the UK could nevertheless exercise some control over Chinese emissions. To 'control' something is roughly the same as having the ability to affect the likelihood of something occurring by undertaking to bring it about, and control thus understood does not presuppose effective jurisdiction. For example, the UK could introduce border carbon adjustments that make imports from China more expensive, which in turn could encourage the Chinese state to pass laws mandating greener production.[16] Call this kind of control, where a country affects the laws and policies of another country through regulating its own consumption, *indirect control*.

The opportunity to exercise indirect control shows that even if we were to consider the UK's responsibility to be to lower emissions emanating from Chinese territory, having jurisdiction is not a precondition for discharging that responsibility. Since the UK can affect emissions emanating from China through demand-side measures such as carbon tariffs, its lack of jurisdiction at most means that its position is *worse* in that regard than that of China. We will return to this point shortly, but for now the conclusion we can draw is that lack of jurisdiction over foreign territories does not render other states unable to affect the emissions emanating from those territories. Hence, we cannot condemn CB accounting because it ascribes prospective responsibilities to states that states cannot discharge.

There is a more serious problem with P2. The premise assumes that CB accounting gives importing countries a responsibility to lower emissions emanating from exporting countries. But this is actually not what CB accounting entails. CB accounting only says that the UK must lower *its own* emissions, *parts* of which are due to consuming goods imported from China. CB accounting is a method for constructing national emissions inventories, so when it ascribes more emissions to a net-importer of GHGs, it does not say that net-importers should be ascribed someone else's emissions. Instead, it offers an alternative approach to calculating which emissions are properly countries' own emissions. This reinforces the point that CB accounting does not ascribe responsibilities that states cannot discharge.

---

such cases, however, we cannot speak of the UK having 'discharged' its prospective responsibility since the success conditions are met independently of what the UK does. The charge, then, is that CB accounting is unfair to the UK in that it assigns a task over which it has no, or little, control, not because the *task as such* is impossible or overly demanding. For 'ought implies can', see Stern (2004).

[16] As Roser and Tomlinson explain, border carbon adjustments are policies that 'focus on applying climate policy to imports and exports and thereby adjusting the differential costs for consumers and producers in countries with different climate policies' (Roser & Tomlinson 2014, p. 228; cf. Böhringer et al. 2012).

After all, if it turns out that the UK is unable to affect the territorial emissions in China, it could choose to reduce its domestic emissions instead or simply consume less. It is only when it is *impossible* to reduce one's emissions sufficiently without reducing foreign emissions that the fairness argument from prospective responsibility succeeds in showing that there might be something unfair about CB accounting.

The fairness argument from prospective responsibility is thus not in good health. Although the argument flows from a plausible normative principle—agents should not be ascribed tasks the success or failure of which is independent of what they do—it does not establish that CB accounting violates this principle. States typically *are* able affect emissions emanating from elsewhere, and CB accounting does not entail that states can only discharge their climatic responsibilities by reducing imported emissions anyway.

4.3 The effectiveness argument

We just noted that it is an exaggeration to say that states cannot affect emissions emanating from other territories since states can exercise 'indirect' control over production taking place abroad. However, this is compatible with saying that states have *less* control over such production compared to the state that has jurisdiction over the relevant territory. The effectiveness argument exploits this point. In this section, I argue that this argument shows that we have a reason to prefer PB accounting to CB accounting in some circumstances.

The effectiveness argument assumes that the accounting method is ultimately about distributing the task of managing and reducing emissions, that is, prospective responsibility. Its normative premise is that the responsibility to perform a task should be ascribed to the agent that is most likely to *discharge* that responsibility.[17] The argument is thus forward-looking. We can state it as follows:

> P1. The responsibility to reduce emissions should be ascribed so that greater emissions reductions are more likely.
>
> P2. Production-based accounting ascribes responsibility to reduce emissions so that greater emissions reductions are more likely.

[17] A further question is whether ascribing prospective responsibility in this way would also be more *efficient*, meaning that tasks are performed at lower cost. For the sake of brevity, I will disregard this aspect, but since economically efficient mitigation probably also means greater emission cuts, we should not overplay the distinction between efficiency and effectiveness here.

Therefore, production-based accounting should be adopted.

Notice that there is a gap in the argument. Since P1 does not say that the responsibility to reduce emissions should *only* be ascribed with an eye to effectiveness, it does not establish the conclusion that PB accounting 'should be adopted.' This is instructive because it shows that even if we were convinced that PB accounting leads to deeper emission cuts, this would not automatically show that this accounting method is preferable. We might care about things other than effectiveness in assigning prospective responsibility such as that it is fair to the burden takers.[18] Still, reducing emissions is highly important. If PB accounting leads to more mitigation than CB accounting, this counts strongly in its favor.

Since P1 is likely true when these qualifications are borne in mind, the relevant question is whether P2 is true. What ground is there for thinking that PB accounting makes greater emissions reductions more likely? This premise is the conclusion of another argument:

P1. Compared to other states, the national state is best placed to affect the emissions emanating from its territory.

P2. Production-based accounting incentivizes states to lower the emissions emanating from their territory.

P3. Therefore, production-based accounting incentivizes the best-placed state to lower emissions emanating from its territory (from 1 and 2).

P4. If we incentivize the best-placed agent to reduce emissions, greater emission reductions are more likely.

Therefore, production-based accounting makes greater emission reductions more likely (from 3 and 4).

The argument is valid and has immediate force. There is admittedly an ambiguity in that the argument aspires to show that PB accounting probably leads to more emission reductions *overall* whereas the premises only speak of emissions emanating *from territories*. However, except for emissions that are not currently attributed to any country, all emissions emanate from some territory. The argument is that if we put emissions on the ledger of the state or government with jurisdiction

[18] For an exploration of the tension between fairness between burden takers and effectiveness, see Caney (2014).

over the territory from which emissions emanate, this makes greater emissions reductions more likely compared to when we count them at the point of final consumption.

Let us go through the argument systematically. Note first that premise P2 seems true. While a state may not care about reducing emissions in the first place, P2 makes the more limited point that if states' climate burdens are sensitive to their level of emissions, and states have an interest in reducing their climate burdens, then PB accounting gives states a reason to care about their territorial emissions. The premise does look problematic if interpreted as saying that *only* PB accounting incentivizes China to lower Chinese territorial emissions. China would have an incentive to lower its territorial emissions under CB accounting, too, since this would enable it to export more goods. But the logic behind P2 is simply that PB accounting incentivizes states to focus on reducing their own territorial emissions. Territorial emissions are the currency through which states live up to their mitigation pledges.

The argument faces objections in relation to the key premises P1 and P4, though. Consider first P4. Even if we concede that the national state is best placed to affect the emissions emanating from a territory, such that PB accounting would incentivize the best-placed agent to reduce those emissions, there is no guarantee that this would lead to emissions reductions, let alone greater reductions. For one thing, some states might simply lack an ambition to reduce their territorial emissions. The fact that such states remain the best placed to reduce their territorial emissions *if they try* is beside the point. For another thing, even if a state does want to reduce its territorial emissions, the extent of these reductions could be smaller than what would have occurred under a system of CB accounting. After all, despite being in a worse position to reduce emissions taking place in a territory, other states could have sufficiently greater climate ambitions, or economic muscle, that the result of demand-side measures is a greater reduction in emissions.

There are, then, some doubts about P4. Still, it seems true enough in a general sense that a problem is more likely to be addressed if we incentivize the agents with the most ability to address a problem to address it. Suppose we could incentivize different agents to perform a task to the same degree. It would then be exceedingly odd if the agent with the greatest ability to perform the task would not generally be the most reliable performer of the task.

The key questions instead pertain to the pivotal premise P1. Unless effective jurisdiction gives states greater control over its territorial emissions, the effectiveness argument breaks down completely. A first thing to note is that P1 says more than just that states have more control over the emissions emanating from their territory than over other emissions. It makes the stronger and more contentious

claim that the national state is better placed to control the emissions emanating from a territory compared to *other states*. We can distinguish between two different comparisons:

> *The narrow effectiveness comparison.* For any state, it is true that the state has more control over its territorial emissions than it has over other emissions.
>
> *The wide effectiveness comparison.* For any territorial emissions, it is true that the national state has more control over these emissions than other states do.

The narrow effectiveness comparison seems plausible enough, but P1 rests on the wide effectiveness comparison. Thus, to assess the truth of P1 we must consider states' avenues for affecting the emissions emanating from other states and then compare the effectiveness of those avenues to the direct control jurisdiction enables.

The case *against* P1 is the following: Under a system of CB accounting, states have an interest in the export sectors of other countries, and to reduce their emissions (which now include imported emissions) they may exercise what I previously called indirect control over them. The clearest example of this is border carbon adjustments, but there are other kinds of demand-side measures, such as product standards or public information campaigns that seek to convince citizens to buy fewer imported goods (Roser and Tomlinson 2014). Such measures are no less likely to lead to emissions reductions than direct regulation of the emissions at their geographical source. Thus, the national state does not have more control over their territorial emissions compared to other states.

The case *for* P1 is that jurisdiction gives states additional tools with which to pursue climate mitigation. In the literature on environmental regulation, it is commonplace to distinguish between policies that harness economic incentives and command-and-control (Baldwin *et al.* 2011).[19] Under incentive-based policies, the regulator seeks to achieve desired outcomes through economic incentives, for example, by offering tax breaks to companies who pollute less. Command-and-control type policies, by contrast, seek to achieve desired outcomes by 'imposing standards backed by sanctions' (Baldwin *et al.* 2011, p. 106). Thus, command-and-control type regulation makes direct use of the law, and may mandate or proscribe a certain behavior. The extra regulatory tools this offers ensure that states are generally best placed to affect the emissions emanating from their territory. Whereas states are at the mercy of indirect economic measures when it comes to

[19] For a more nuanced taxonomy over different kinds of regulation, see Coria and Sterner (2011).

reducing emissions taking place elsewhere, they can directly intervene in what goes on in their own backyard.

I do not know which case is more plausible, and I suspect that adjudicating the issue is enormously complicated. But let me add a few observations that may help clarify the issue somewhat, before I turn to how I think we should think of the effectiveness argument as a whole. First, the claim that only PB accounting gives states access to direct jurisdictional control over all their emissions is probably correct. While we could imagine states that legislate imported consumption just as much as they legislate production, in practice this is unlikely. In our open economies, it is more common for states to impose laws on factories than on individual consumers when it comes to reducing greenhouse gas emissions.

Second, we can endorse P1 while maintaining that regulation based on economic incentives is generally more effective than command-and-control. Many social scientists argue that incentives-based regulation is better in virtue of its greater flexibility, lower regulatory burden, and lesser risk of regulatory capture.[20] But conceding this point does not mean that command-and-control regulation is unimportant. Command-and-control might be necessary when incentive-based policies prove inadequate to achieve the desired outcome.

However, third, even if we grant that PB accounting gives states access to regulatory tools which would be needed *if* indirect policies fail to bring about the desired outcomes, this does not show that these tools *are* needed for the simple reason that indirect policies might *actually* be enough to bring about the desired outcomes. Put differently, if a carbon tariff implemented by importing states is enough to reduce emissions in China, the fact that direct legislative action by the Chinese state would have been needed to reduce emissions if the tariff did not work is simply irrelevant. This is important because if the additional control jurisdiction offers is never needed, it can give us no reason to prefer PB accounting. (Remember, the idea we are now discussing is that PB accounting is effective because it ensures that states seek to reduce the emissions over which they have the most control.)

There is admittedly some reason to doubt the adequacy of indirect or demand-side policies. Policies that aim to change consumption behavior through labelling and information campaigns, for example, tend to be relatively ineffective (Peters et al. 2016). Economic policies such as carbon tariffs—if they are even legal according to current international trade rules—face the problem of price inelasticity, and in addition risk to merely divert exports to another importer as opposed to reducing the territorial emissions of the exporter (Roser and Tomlinson 2014).[21] However, as

[20] A well-known statement of this view is Tietenberg (1991). For summaries of the debate, see Goulder and Parry (2008); Baldwin et al. (2011); and Coria and Sterner (2011).

[21] Price inelasticity is a problem for PB accounting too, since a state will be unable to reduce their

evidenced by the large and inconclusive debate on the environmental effects of consumption-based emissions policies, these problems should not be seen as conclusive reason to think that PB accounting is more effective. It is simply unclear at this point whether a move to CB accounting would make greater emissions reductions more likely.[22]

Fourth and most importantly, even if the ability to regulate emissions directly at the source means that PB accounting *in general* ascribes prospective responsibility to the agent that is best placed to control emissions, this does not mean that it *always* does so. To give just one example, it might be that carbon tariffs employed by the EU (setting their legality to one side) could do more to reduce global emissions than assigning unambitious states the responsibility to reduce their territorial emissions. In all such cases, the effectiveness argument from jurisdictional control is unsuccessful — it gives us no reason to prefer PB accounting to CB accounting. Thus, the truth of P1 is probably something we must assess on a case-to-case basis.

This means that the effectiveness argument merely establishes an important *potential* ground for preferring PB accounting. The soundness of the argument must be established inductively. More specifically, we must assess the extent to which the national state or government is best placed to affect the territorial emissions in each case, and then tally up the results into an aggregate picture. If it turns out that the national state or government is best placed in most cases, or in relation to most emissions, then that is a reason to stick with PB accounting. But if it turns out that other states are better placed, then that is a reason to switch to CB accounting instead.

Though tallying up individual cases in this way appears to be the right approach, this is not to suggest that it would be easy to *determine* whether the national state or government is best placed in each individual case, and I shall certainly not attempt to do so. My point is simply that there is a coherent case for thinking that PB accounting might be preferable to CB accounting on effectiveness grounds. Finding out whether this case holds should be a priority for economists and political scientists interested in international climate policy.

Three further points are worth making. First, what happens once we factor in

territorial emissions by introducing, say, a carbon tax if domestic consumers are insensitive to price increases for domestically produced goods. But the point is that the state has access to command-and-control type regulation in the territory over which it has jurisdiction, and such regulation is not (as) plagued by price inelasticity.

[22] For the studies that seek to model the effects of consumption-based emissions policies, see, e.g., Steckel *et al.* (2010); Jakob *et al.* (2014); Springmann (2014); Steininger *et al.* (2014); and Lininger (2015). Peters et al. (2016, p. 52) argue that the border carbon adjustment literature suggests that such indirect measures reduce policy-induced carbon leakage, but only marginally so. It is worth pointing out that these studies are based on modelling since actual consumption-based emissions policies are yet to be tried out.

uneven political will? Proponents of CB accounting frequently stress that this method would be more effective only because it would attribute more emissions to countries that have more progressive climate ambitions.[23] In the current configuration of the world, where net importers tend to be rich countries with more stringent mitigation targets, CB accounting could indeed ramp up the global mitigation effort compared to the status quo for this reason. However, that possibility does not invalidate the point that PB accounting might have an advantage over CB accounting when employed among states with similar climate ambitions. Needless to say, if political will is lacking all around, it does not matter how we count emissions.

Second, effectiveness is not all we should care about when choosing accounting method. Suppose that we would get greater emission reductions by counting emissions at their geographical point of origin, but that this would also release rich countries from prospective responsibilities they would have under CB accounting. If we subscribe to a view of climate justice according to which rich countries should cover the costs of climate policy, the latter would be a reason in itself to regard PB accounting as morally worse. Distributive justice in burden sharing bears on the choice of accounting method quite independently of effectiveness (redacted 1).

Third, even if we focus on effectiveness, we must distinguish between the environmental effects of an accounting method and the environmental effects of *switching* to that method. Some warn that attempting to switch to CB accounting would slow down the urgently needed international action on climate change (Peters et al. 2016; Afionis et al. 2017). If this is right, another effectiveness-based argument in favor of sticking with PB accounting is simply that it is the system we currently have.

7. Conclusion

I investigated the significance of jurisdiction for the choice of accounting method of greenhouse gases. Making use of the distinction between retrospective and prospective responsibility, I assessed three different arguments from jurisdiction against consumption-based accounting. I argued that the effectiveness provides a strong potential reason against consumption-based accounting. To the extent jurisdictional control is needed to reduce some emissions, and production-based accounting incentivizes states to reduce these emissions, there is a reason of environmental effectiveness for sticking with production-based accounting.

[23] Another important reason, relevant especially for border carbon adjustments, is that pricing imports in accordance with the domestic carbon price protects the domestic industry, ensuring that states with more ambitious climate targets are not put at a disadvantage (Roser and Tomlinson 2014).

References

Afionis, S., et al. (2017). Consumption-based carbon accounting: does it have a future? *WIREs Climate Change,* 8(1), 1–19.

Baldwin, R. et al. (2012). *Understanding Regulation: Theory, Strategy, and Practice*. Oxford: Oxford University Press.

Barrett, J., et al. (2013). Consumption-based GHG emissions accounting: a UK Case Study. *Climate Policy*, 13(4): 451-470.

Böhringer, C., et al. (2012). The Role of Carbon-Adjustments in Unilateral Climate Policy: Overview of an Energy Modelling Study. *Energy Economics,* 34(2), 97–110.

Caney, S. (2005). Cosmopolitan justice, responsibility, and global climate change. *Leiden journal of international law,* 18(4), 747–775.

Caney, S. (2014). Two Kinds of Climate Justice. *The Journal of Political Philosophy,* 22(2), 125–149.

Coria, J. and Sterner, T. (2011). Natural resource management: challenges and policy options. *Annual Review of Resource Economics*, 3(1): 203–230.

Davis, S. and Caldeira, K. (2010). Consumption-based accounting of CO_2 emissions. *PNAS*, 107(12), 5687–5692.

de Sadeleer, N. (2002). *Environmental Principles: From Political Slogans to Legal Rules*. Oxford: Oxford University Press.

Duus-Otterström, G. and Hjorthen, F. (2018). Consumption-based emissions accounting: the normative debate. *Environmental Politics*. Online first.

Goodin, R. (1985). *Protecting the Vulnerable.* Chicago: The University of Chicago Press.

Goulder, L. and Parry, I. (2008). Instrument choice in environmental policy. *Review of Environmental Economics and Policy*, 2(2): 152–174.

Grasso, M. (2017). Achieving the Paris Goals: Consumption-based carbon accounting. *Geoforum*, 79, 93–96.

Grasso, M. and Roberts, T. (2014). A compromise to break the climate impasse. *Nature Climate Change*, 4: 543–549.

Hussain, A. (2012). Is ethical consumerism an impermissible form of vigilantism? *Philosophy & Public Affairs*, 40, 111–143.

IPCC (2006). *2006 IPCC Guidelines for National Greenhouse Gas Inventories*. Kanagawa: IGES.

Jakob, M., et al. (2014). Consumption versus production-based emission policies. *Annual Review of Resource Economics*, 6: 297–318.

Mackie, J.L. (1965). Causes and conditions. *American Philosophical Quarterly*, 2, 245–264.

Miller, D. (2012). Territorial rights: concept and justification. *Political Studies*, 60, 252–268.

Lawford-Smith, H. (2015). Unethical consumption and obligations to signal. *Ethics & International Affaris*, 29, 315–330.

Lenzen, M., et al. (2007). Shared producer and consumer responsibility – theory and practice. *Ecological Economics,* 61(1), 27–42.

Lininger, C. (2015) *Consumption-Based Approaches in International Climate Policy*. Dordrecht: Springer.

Liu L. (2015). A critical examination of the consumption-based accounting approach: has the blaming of consumers gone too far? *WIREs Clim Change*, 6:1–8.

Miller, D. (2007). *National Responsbility and Global Justice*. Oxford: Oxford Universoity Press.

Peters, G.P. (2008). From production-based to consumption-based national emission inventories. *Ecological Economics,* 65(1), 13-23.

Peters, G.P. and Hertwich, E.G. (2008). Post-Kyoto greenhouse gas inventories: production versus consumption. *Climatic Change,* 86(1), 51–66.

Peters, G.P., et al., (2016). Global Environmental Footprints. Norden. Avalibale at: http://folk.uio.no/roberan/docs/Peters2016_NordicFootprints.pdf (Accessed 2018-11-09).

Roser, D. and Tomlinson, L. (2014). Trade policies and climate change: border carbon adjustments as a tool for a just global carbonrRegime. *Ancilla Iuris,* 75, 223–245.

Roser, D. and Seidel, C. (2017). *Climate Justice*. London: Routledge.

Scanlon, T.M. (1999). *What We Owe to Each Other*. Cambridge: Harvard University Press.

Shue, H. (1993). Subsistence emissions and luxury emissions. *Law & Policy*, 15, 39-59.

Springmann, M. (2014). Integrating emissions transfers into policymaking. *Nature Climate Change*, 4, 177–181.

Steckel, J., et al. (2010). Should carbon-exporting countries strive for consumption-based accounting in a global cap-and-trade regime? *Climate Change*, 100: 779–786.

Steininger, K., et al. (2014). Justice and cost effectiveness of consumption-based versus production-based Approaches in the Case of Unilateral Climate Policies. *Global environmental change,* 24, 75–87.

Steininger, K., et al. (2016). Multiple carbon accounting to support just and effective climate policies. *Nature Climate Change*, 6(1), 35–41.

Stern, N. (2004). Does 'ought' imply 'can'? And Did Kant think it does? *Utilitas* 16: 42–61.

Tietenberg, T.H. (199)1. Economic instruments for environmental regulation. *Oxford Review of Economic Policy*, 6(1): 17–33.

Vallentyne, P. (2008). Brute luck and responsibility. *Politics, Philosophy & Economics,* 7(1), 57–80.

Valverde, M. (2014). Studying the governance of crime and security. *Criminology and Criminal Justice*, 14, 379–391.

van der Poel, I. (2010). The relation between forward-looking and backward-looking responsibility. In: Vincent, N. et al. (Eds.) *Moral Responsibility: Beyond Free Will and Determinism*. Dordrecht: Springer, 37–51.

Vincent, N. (2010). A structured taxonomy of responsibility concepts. In: Vincent, N. et al. (Eds.) *Moral Responsibility: Beyond Free Will and Determinism*. Dordrecht: Springer, 15–35.

Young, I.M. (2006). Responsibility and global justice: a social connection model. *Social Philosophy and Policy*, 23, 102–130.

Paul Bowman[1]

On the Alleged Insufficiency of the Polluter Pays Principle

Several theorists of climate change justice have argued that the polluter-pays principle fails to assign duties that, if fulfilled, would be sufficient to prevent or compensate for all climate change-induced harm to persons. This paper contends that their argument for this claim rests on a faulty account of how the costs of rectifying a collectively-caused harm or threat of harm should be allocated among agents who have incurred duties of corrective justice to bear these costs. Given a more plausible account of how these costs should be allocated, it is likely that the polluter-pays principle does in fact assign duties that, if fulfilled, would be sufficient to prevent or compensate for all climate change-induced harm to persons.

[1] Institute for Futures Studies, paul.bowman@iffs.se. Financial support by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences) is gratefully acknowledged.

I. Introduction

The polluter pays principle holds, roughly, that those who have produced excessive amounts of greenhouse gas emissions have duties to bear the costs of addressing harmful climate change. These costs include the costs required to prevent further anthropogenic changes to the climate system (mitigation costs), the costs required to prevent climate change impacts from harming persons (adaptation costs), and the costs required to compensate persons who have been harmed by climate change impacts (compensation costs).

The polluter pays principle, as it is typically understood by theorists of climate change justice, is based on a widely-endorsed principle of corrective justice: roughly, the principle that agents who wrongfully contribute to a harm or threat of harm incur a duty of corrective justice to bear the costs of rectifying the harm or threat (by preventing, mitigating, or compensating for it). Because climate change has resulted in harm and threatens significant future harm, and because it is plausible that many agents have wrongfully contributed to climate change by producing excessive amounts of greenhouse gas emissions, many theorists have found the polluter pays principle to be an attractive principle for assigning duties to bear the costs of addressing harmful climate change.[2]

While many theorists accept some version of the polluter pays principle, there appears to be widespread agreement among theorists of climate change justice that the principle is inherently limited or incomplete in one particular respect. Specifically, these theorists accept:

[2] Theorists who endorse the polluter pays principle include, e.g., Shue (1999), Caney (2005; 2010), Vanderheiden (2008), Bell (2010), Roser and Meyer (2010), Cripps (2013), and Zelletin (2014). Different theorists endorse different versions of the principle. Versions of the principle tend to differ according to the kinds of greenhouse gas emissions the production of which triggers a duty to bear costs. For example, some theorists have defended versions of the principle according to which an agent only incurs a duty to bear costs by producing emissions that exceed the agent's fair share of global 'safe' emissions (e.g., Bell, 2010; Cripps, 2013). Others have defended versions of the principle according to which an agent only incurs a duty to bear costs by producing emissions after the date at which the existence of human-caused climate change became reasonably well-established (e.g., Caney, 2005; Bell, 2010).
Some theorists hold that the polluter pays principle only assigns duties to bear adaptation and compensation costs, whereas different principles are needed to assign duties to bear mitigation costs (e.g., Vanderheiden, 2008; Roser and Meyer, 2010). Although I cannot here explain why, I do not find it plausible to restrict the principle in this way. I assume that the polluter pays principle assigns duties to bear mitigation costs as well as adaptation and compensation costs. My argument in what follows is aimed primarily at the views of those who share this assumption, though my argument could be suitably reformulated to address defenders of the more restricted version of the principle.

> *The Insufficiency Claim*: the total costs that the polluter pays principle assigns agents duties to bear are not sufficient to cover the costs required to prevent or compensate for all climate change-induced harm to persons.

In other words, the Insufficiency Claim holds that if we were to add up all the costs that the polluter pays principle assigns agents duties to bear, these costs would be less than the total costs needed to prevent or compensate for all climate change-induced harm to persons that has occurred and that is threatened to occur. According to the Insufficiency Claim, then, even if all agents fulfill their duties assigned by the polluter pays principle, and if no agents bear additional preventative or compensatory costs, then some people will suffer uncompensated climate change-induced harm.

The Insufficiency Claim is significant because it is widely-accepted that a morally satisfactory response to climate change requires that persons are protected from all climate change-induced harm, or if protection is not possible or morally feasible, then those persons who have been harmed (or will be harmed) by climate change are compensated (e.g., Caney, 2010; Baatz, 2013). Thus, it is widely-accepted that a complete account of climate change justice must include principles that together assign duties that, if fulfilled, would be sufficient to cover the costs required to prevent or compensate for all climate change-induced harm to persons. So, if the Insufficiency Claim is true, then principles in addition to the polluter pays principle are needed to assign duties to bear the remaining costs required to prevent or compensate for all climate change-induced harm to persons. Theorists have typically identified these principles as those that assign duties to bear costs on the basis of agents' ability to pay (e.g., Caney, 2005; 2010) or on the basis of the benefits that agents have received from activities that produced greenhouse gas emissions (e.g., Baatz, 2013; Duus-Otterström, 2014).

In what follows, I will evaluate the Insufficiency Claim. I will contend that the primary argument that theorists have advanced in support of the claim rests on a faulty account of how the costs of fully rectifying a collectively-caused harm or threat should be allocated among those who incur duties of corrective justice to bear at least some of these costs. Moreover, I will argue that given a more plausible account of how these costs should be allocated, it is likely that the Insufficiency Claim is false. At the very least, I hope to show that the question is more complicated than theorists have hitherto assumed.

The basic plan for the paper is this. In section II, I will present what I take to be the main argument for the Insufficiency Claim. In sections III through V, I will evaluate the account that underlies the central premise in that argument. I will argue that this account is false, and I will defend a competing account. In Section VI,

I will consider the implications of my account for the Insufficiency Claim. I will argue that the Insufficiency Claim is likely false. I will conclude briefly in Section VII.

II. The main argument for the Insufficiency Claim

As I have said, the Insufficiency Claim has received widespread support among theorists of climate change justice. What arguments, then, do defenders of the Insufficiency Claim provide in support of it?

According to its defenders, the Insufficiency Claim is supported, first, by the fact that the polluter pays principle does not, for obvious reasons, assign duties to agents who have emitted greenhouse gases but who are now dead. In one of his several articles that discusses the problem that dead emitters pose for the polluter pays principle, Simon Caney (2005) writes that:

> One problem with applying the 'polluter pays' principle to climate change is that much of the damage to the climate was caused by the policies of earlier generations. It is, for example, widely recognized that there have been high levels of carbon dioxide emissions for the last two hundred years, dating back to the Industrial Revolution in western Europe. This poses a simple, if also difficult, problem for the 'polluter pays' principle: who pays when the polluter is no longer alive? (p. 756)

Caney argues, ultimately, that the polluter pays principle, "cannot say who should bear the costs of climate change caused by past generations," and so a different principle is needed to assign duties to bear these costs (2005: 761).[3]

Similarly, Göran Duus-Otterström (2014) writes:

> [The polluter pays principle], however, cannot handle emissions that were emitted by agents who are no longer in existence—what I will call *past emissions*. Even though such emissions may continue to force the climate, there is no way of making the polluters responsible for them to 'pay.' Only the living—by definition not responsible for having caused past emissions—can shoulder duties. Thus, if we now want to handle or compensate for past emissions, some will have to bear *extra burdens*. (p. 449, his emphasis)

[3] See Caney's 2010 for a similar argument. In both his 2005 and his 2010, Caney argues that the polluter pays principle should be supplemented with a version of the ability to pay principle.

Like Caney, Duus-Otterström argues that duties to bear the burdens not assigned by the polluter pays principle must be assigned according to a different principle.[4]

Here, finally, is Moellendorf (2014):

> [A] serious practical problem of insufficient collection exists [. . .] because many of the damage-causing stock of atmospheric CO_2 was produced by people who are now long dead. Despite the remarkable advances of modern medicine, there is little prospect of recouping costs from them. So, any assignment of responsibility on the basis of [the polluter pays principle] necessarily ensures that full costs will not be recouped. (p. 166)

According to Caney, Duus-Otterström, and Moellendorf, although many agents who are now dead produced greenhouse gas emissions, the polluter pays principle does not assign duties to bear the costs of addressing harmful climate change on the basis of these emissions. Therefore, duties to bear the costs attributable to these emissions must be allocated according to some other principle or principles (or left unaddressed).[5]

In addition to not assigning duties on the basis of emissions of those who are now dead, some versions of the polluter pays principle do not assign duties on the basis of some currently living agents' greenhouse gas emissions. Some theorists who endorse one of these versions of the polluter pays principle have pointed to this fact as further evidence for the Insufficiency Claim.

For example, Baatz (2013) argues for a version of the polluter pays principle that does not assign duties on the basis of emissions agents produced prior to 1990, given that prior to 1990, agents were non-culpably ignorant that their emissions were

[4] According to Duus-Otterström (2014), this principle is what he calls the "inherited debt principle," which holds, roughly, that extra burdens should be borne by those who have inherited holdings that resulted from uncompensated overuse of the atmosphere.

[5] Also see Page (2008), who states that "the 'contribution to problem' approach cannot deal well with the fact that many of the individual citizens and policymakers responsible for their countries' historical emissions are now dead" (p. 559).

It is important to note that, according to supporters of the Insufficiency Claim, expanding the *kind* of agent to which the polluter pays principle assigns duties to include collective entities like states and corporations does not eliminate the problem of dead emitters, for three reasons. First, with respect to states, Caney (2010) argues that it would be unfair for the current citizens of a state to bear the costs attributable to emissions generated long before these citizens were alive. If this is correct, then it appears that states can incur duties only on the basis of emissions they produced in the more recent past. Second, Caney (2010) notes that some of the states and corporations that produced emissions no longer exist. Therefore, the polluter pays principle does not assign duties to these "dead" states and corporations. And third, even if states and corporations can be held liable on the basis of the emissions they produced in the distant past, many now dead individual persons *also* produced emissions. Thus, the problem remains: many dead agents emitted greenhouse gases, but the polluter pays principle does not assign duties to bear costs on the basis of their emissions.

contributing to harmful climate change. Baatz states that an implication of the principle is that the principle cannot address "who has to pay for damages caused by pre-1990 emissions" (p. 96).[6] Similarly, Caney (2010) argues for a version of the polluter pays principle that does not assign duties on the basis of emissions produced by activities that are necessary for one's "fundamental survival" (p. 213). According to Caney, such emissions are legitimate, or morally justified, and thus do not trigger duties of corrective justice. Consequently, Caney argues that the polluter pays principle must be supplemented with principles needed to "deal with [. . .] harmful climate changes that stems from [. . .] the (legitimate) emissions of the disadvantaged" (p. 213).[7]

We can, I believe, capture the basic form of the argument for the Insufficiency Claim as follows:

> (P1) The polluter pays principle does not assign duties to bear costs on the basis of a significant portion of greenhouse gas emissions.
>
> (P2) If the polluter pays principle does not assign duties to bear costs on the basis of a significant portion of greenhouse gas emissions, then the total costs that the polluter pays principle assigns agents duties to bear are less than the total costs required to prevent or compensate for all climate change-induced harm to persons.
>
> *Insufficiency Claim*: The total costs that the polluter pays principle assigns agents duties to bear are less than the total costs required to prevent or compensate for all climate change-induced harm to persons.

Call this the "Main Argument."

The Main Argument appears to be valid. Moreover, (P1) is clearly true. Defenders of the Insufficiency Claim have identified three kinds of emission on the basis of which at least some versions of the polluter pays principle do not assign duties to bear costs:

---

[6] See Caney (2005) and Bell (2010) for similar arguments.

[7] Caney (2010) claims that, in addition to the emissions of those now dead and emissions that are morally justified, the polluter pays principle cannot deal with climate change that stems from natural factors. Given space constraints, I have chosen not to address this argument here. I plan to address the topic elsewhere. In short, I think that we should treat natural factors as background conditions for the purposes of determining the magnitude of anthropogenic climate change-induced harms that the polluter pays principle assigns duties to address. Since non-anthropogenic climate change accounts for only a very small amount of observed warming, it is unlikely that non-anthropogenic climate change would, in the absence of anthropogenic activities, result in any serious harm. Therefore, the existence of natural climate change provides little or no support for the Insufficiency Claim.

(a) Emissions produced by those who are now dead ("*past emissions*")

(b) Emissions produced by activities that are morally justified, including activities required for survival ("*justified emissions*"), and

(c) Emissions produced by agents who have a full excuse for producing them, including agents who were non-culpably ignorant that their emissions were contributing to harmful climate change ("*excused emissions*").

Note that no version of the polluter pays principle assigns duties on the basis of past emissions, whereas some, but not all, versions of the principle assign duties on the basis of agents' justified emissions or excused emissions. Going forward, I shall assume that the correct version of the polluter pays principle does not assign duties on the basis of emissions from any of the above categories. Henceforth (and for lack of a better term), I shall refer to the emissions from any of these categories as *non-liable emissions*, or NLEs.

So (P1) is true. What about (P2)? (P2) states that an implication of (P1) is the Insufficiency Claim: that the total costs that the polluter pays principle assigns agents duties to bear are less than the total costs that are required to prevent or compensate for all climate change-induced harm. Specifically, defenders of the Insufficiency Claim hold that the polluter pays principle fails to assign duties to bear the costs of preventing or compensating for climate change-induced harm that "stems from" or is in some way attributable to NLEs.

Although (P2) is the crux of the Main Argument, proponents of the Insufficiency Claim have not, to my knowledge, explicitly defended it. As the passages I quoted above indicate, these theorists proceed directly from the fact that NLEs exist to the Insufficiency Claim. In the remainder of the paper, I shall attempt to determine whether the existence of NLEs implies or in any way supports the Insufficiency Claim. My strategy will be to investigate different accounts of how the principle of corrective justice on which the polluter pays principle is based should assign agents duties to bear the costs of rectifying a collectively-caused harm or threat when the actions of those who are now dead, as well as justified actions and excused actions, have contributed to the harm or threat. I will argue that the account that underlies (P2) is false. I will also argue for a competing account, which neither implies nor supports the Insufficiency Claim.

III. Corrective Justice and the share of the total account

The moral principle that underlies the polluter pays principle is roughly the following:

> *Principle of Corrective Justice–Multiple Agents* (CJ-MA): If several agents' actions contribute to some harm or threat of harm h, then an agent A incurs a duty of corrective justice to bear some of the costs of rectifying h (preventing, mitigating, or compensating for h) just in case A makes an unjustified and unexcused contribution to h.

Although (CJ-MA) specifies the conditions under which an agent incurs a duty of corrective justice, the principle does not address the extent of an agent's duty, or the costs that the agent incurs a duty to bear in service of rectifying the harm or threat.

Theorists who have discussed the polluter pays principle have typically stated or suggested that the extent of an agent's duty assigned by the principle must reflect, or be in proportion to, the amount of (unjustified and unexcused) emissions that the agent produced. Caney (2005), for example, states that the polluter pays principle is based on the principle that "if actors X, Y, and Z perform actions which together cause pollution, then they should pay for the cost of the ensuing pollution in proportion to the amount of pollution that they have caused" (p. 754).

Therefore, I will assume that (CJ-MA) also includes what I will call the *proportionality constraint*:

> the extent of the costs of rectifying h that A incurs a duty to bear must be in proportion to the extent of A's unjustified and unexcused contributions to h.

There are different ways in which one might interpret the proportionality constraint. Moreover, the correct interpretation of the constraint is relevant to whether the Insufficiency Claim is true. As I explain below, theorists have tended to assume (unjustifiably, in my view) an interpretation of the proportionality constraint that has led these theorists to accept (P2) and thus the Insufficiency Claim. For now, it is enough to note that the proportionality constraint implies that the extent of an agent's unjustified and unexcused contributions to a harm or threat is one factor that determines the extent of the costs that the agent incurs a duty of corrective justice to bear. This leaves open whether there are other factors that determine the extent of an agent's duty.

The following example will help illustrate (CJ-MA). I will use this example (as well as several variations) in my subsequent discussion of the principle:

> *Waste:* Each of four agents—A, B, C, and D—independently dumps ten gallons of the same kind of industrial waste into a creek that feeds into a victim V's water supply. Each agent dumps her waste to avoid a moderate yet affordable disposal fee. V drinks the waste-laced water and goes blind. At the time of action, each agent knows for certain that the other agents either have dumped or will dump their waste. Each agent also knows for certain that her waste, combined with the others' waste, will result in V's going blind, and that V would not go blind (or be harmed at all) if no agent dumps her waste. The total cost of restoring V's eyesight is $1,200, and there are no additional, compensatory costs. A, B, C and D are each capable of fulfilling whatever duty is assigned to her.

I assume that the cost of fully rectifying the harm to V is $1,200. (Henceforth, I will sometimes refer to the cost of fully rectifying some harm or threat as the *magnitude* of that harm or threat.)

Additionally, I assume that none of the agents' contributions are either justified or excused. I assume also that nothing morally differentiates the agents' actions from one another. I assume, finally, that (CJ-MA) implies that each A, B, C, and D incurs a duty to bear one-fourth of the $1,200 cost, or $300. In this case, the total cost that (CJ-MA) assigns agents duties to bear is equal to the total cost of fully rectifying the harm to V.

The implications of (CJ-MA) appear to be relatively straightforward in cases like *Waste* in which all of the contributions to the harm or threat are those on the basis of which the principle assigns duties.[8] However, (CJ-MA) does not, by itself, tell us how the costs of rectifying a collectively-caused harm or threat should be allocated among agents when one or more contributions are *not* those on the basis of which (CJ-MA) assigns duties—i.e., the contributions of those who are now dead, or justified or excused contributions. (I will henceforth refer to such contributions as non-liable *contributions*, or NLCs.)

Let's therefore consider how (CJ-MA) should handle NLCs. Consider the following variation of *Waste:*

[8] This is not to say that there may not be other problems with (CJ-MA) or applying (CJ-MA) to particular cases (e.g., cases in which assessing the extent of an agent's contribution is not straightforward).

> *Dead Polluter:* Same as *Waste*, but A dies before she fulfills her duty to contribute $300. A leaves no estate from which her duty can be fulfilled.

Notice that once A dies, (CJ-MA) does not assign a duty on the basis of A's contribution to the harm, given that A, the would-be duty-bearer, is dead. A's contribution is an NLC.

As before, the magnitude of the harm is $1,200. What, then, is the cost that each of the three remaining agents—B, C, and D—has a duty to bear, given that A's contribution is an NLC? Does each simply have a duty to bear the cost that she would have had to bear had A *not* died (i.e., $300)? This would mean that (CJ-MA) does not assign a duty to bear the remaining $300 of the $1,200 cost. Or does each of the remaining agents incur an additional duty to bear a portion of the cost that A had a duty to bear while she was alive, perhaps in proportion to the extent to which each remaining agent contributes to the harm (i.e., $1,200 divided by three, or $400 each)? Or is there perhaps another way of allocating the costs?

Proponents of the Insufficiency Claim have not, to my knowledge, explicitly argued for an answer to this question.[9] It is highly plausible, however, that they have simply presupposed the truth of what I will call the *Share of the Total Account*.[10] roughly, the view that the cost that an agent P has a duty to bear in service of rectifying a collectively-caused harm or threat h is n, where:

$$\frac{n}{\text{total costs of rectifying h}} = \frac{\text{the extent of P's unjustified and unexcused contributions to h}}{\text{the extent of all contributions to h (including NLCs)}}$$

Consider again *Dead Polluter*. Because B, C, and D each contributes one-fourth of the total waste that results in the harm to V, according to the Share of the Total Account, each must bear one-fourth of the total costs of rectifying the harm to V (one-fourth of $1,200, or $300 each). Thus, (CJ-MA) assigns duties to bear costs totaling $900 of the $1,200 total cost of rectifying the harm. If the Share of the Total Account is correct, then (CJ-MA) does not assign duties to bear the remaining $300 cost.

Notice that if the Share of the Total Account is correct, then (P2) is true: the existence of NLEs implies that the total costs that the polluter pays principle assigns

[9] Though as I shall discuss in section V, Caney (2005) does briefly argue against one answer to this question.

[10] I borrow this name from Parfit (1984). The Share of the Total Account that Parfit discusses resembles but is not identical to the account I discuss here.

duties to bear is less than the total costs required to prevent or compensate for all climate change-induced harm. Therefore, the Share of the Total Account supports the Insufficiency Claim.

I suspect that theorists who endorse the Insufficiency Claim simply presuppose the truth of the Share of the Total Account. There are three reasons for my suspicion. First, the Share of the Total Account provides a straightforward and intuitive way to understand the proportionality constraint. Second, the Share of the Total Account provides unambiguous support for (P2). As I mentioned above, theorists who endorse the Insufficiency Claim do not explicitly defend (P2). If these theorists simply presuppose the truth of the Share of the Total Account, it is understandable why they do not think it is necessary to defend (P2). And third, the Share of the Total Account makes some sense of the idea the polluter pays principle does not assign duties to bear costs that "stem from" NLEs. In *Dead Polluter*, for example, the Share of the Total Account makes sense of the idea that the remaining $300 of the $1,200 total cost stems from A's contribution—namely, the cost stems from A's contribution in the sense that the extent of the cost is strictly proportional to the extent of A's contribution.

Over the next two sections, I will argue that the Share of the Total Account is false. I will also argue for an alternative account of how (CJ-MA) should assign duties to bear the costs of rectifying a collectively-caused harm or threat when some of the factors that contributed to it are NLCs.

IV. Two bad arguments for the Share of the Total Account

There are two arguments for the Share of the Total Account that are worth considering but that can be dispensed with relatively quickly. The first argument seeks to establish the Share of the Total Account on the basis of a metaphysical claim concerning the attribution of certain consequences to a person's action. Perhaps it is *prima facie* plausible that, as a purely metaphysical claim, if an agent contributes n% of all contributions that result in a harm or threat, then the agent individually causes n% of that harm or threat. If this claim is true, and if we also assume that an agent incurs a duty of corrective justice to bear only the cost of rectifying a harm or threat she individually causes, then an agent who contributes n% of all contributions that result in a harm or threat incurs a duty to bear, at most, n% of the cost of rectifying the harm or threat, as the Share of the Total Account states.

However, it is false that if an agent contributes n% of all contributions that result in a harm or threat, then the agent individually causes n% of that harm or threat. To see why, consider two versions of *Waste*. In the first version, suppose that V would

have gone blind if at least thirty-five gallons of waste were dumped in the creek, such that agent A's contribution is necessary, but not sufficient, for V's blindness. In the second version, suppose that V would have gone blind if only twenty-five gallons of waste were dumped in the creek, such that A's contribution is neither necessary nor sufficient for V's blindness. In neither version of the case does it seem that can we, in any meaningful, metaphysical sense, attribute one-fourth of the harm V suffers to A's contribution. In the first version, given that A's contribution is necessary for the harm, it seems that A individually causes the entire harm.[11] In the second version, A's contribution is one of several factors that together cause the entire harm, though the contribution does not seem individually to cause any harm.

There may be some circumstances where by contributing n% of all contributions that result in a harm, the agent individually causes n% of that harm. Suppose, for example, that in *Waste*, the agents dump their ten gallons sequentially, and V's vision gets incrementally worse by an equally bad amount for each ten gallons of waste that is dumped until V is completely blind after the last agent dumps her waste. In this version, perhaps it makes sense to say that A individually causes one-fourth of the harm to V, given that A's contribution is both necessary and sufficient for a loss that is a component of an overall much greater loss. But importantly, most of the harms that have resulted and will result from climate change are not like this. One person's emissions are not, nor will they ever be, necessary and sufficient for a part of any person's climate change-induced loss of life, health, or property in the way that A's dumping her waste is necessary and sufficient for part of V's blindness. Rather, one person's emissions are more like A's contributions in the variations of *Waste* discussed in the previous paragraph. Either a person's emissions are necessary but not sufficient for some climate change-induced harms, or (far more commonly) the emissions are neither necessary nor sufficient for the climate change-induced harms to which the emissions contribute.

Let's turn to the second argument for the Share of the Total Account. Several theorists have argued that an agent's emitting greenhouse gases individually increases *expected* climate change-induced harm by increasing the probability that additional or worse climate change impacts will occur.[12] Moreover, given that there is a roughly linear relationship between total global emissions and total expected climate change-induced harm, these theorists argue that it is possible to approximate the extent to which an agent's emissions individually increase expected harm by dividing her emissions by total global emissions and then by multiplying this

[11] Of course, each of B, C, and D also individually causes the harm. I do not think it is implausible that more than one agent can individually cause some harm.
[12] See, e.g., Hiller (2011), Broome (2012), and Morgan-Knapp and Goodman (2015).

number by the total harm expected to result from total global emissions.[13] If we suppose that each agent incurs a duty to bear only the costs that correspond to the extent to which the agent's unjustified and unexcused contributions individually increase expected harm, then perhaps the Share of the Total Account is correct in virtue of best approximating these costs.

In response, it is false that each agent incurs a duty to bear only the costs that correspond to the extent to which the agent's unjustified and unexcused contributions individually increase expected harm. Consider:

> *Overdetermined Threat:* Same as *Waste*, but at the time of acting, each agent knows for certain that (a) if twenty-five or fewer total gallons of waste are dumped, V will not suffer any harm at all, but (b) if greater than twenty-five total gallons of waste are dumped, V will go blind in the near future, unless significant preventative action is taken (e.g., draining and replenishing the water supply). Each agent also knows for certain that V is not at risk of suffering any harm worse than blindness regardless of how much total waste is dumped (and that no other person is at risk of suffering harm). Finally, each agent knows for certain that the other agents have dumped or will dump their waste, and each agent knows that she cannot prevent any other agent from dumping her waste.

None of the agent's contributions individually increase expected harm. Yet it is highly plausible both that each agent lacks either justification or excuse for her contribution, and that each agent incurs a duty of corrective justice to bear one-fourth of the costs of preventing the harm they together threaten. Therefore, if the Share of the Total Account is correct, it is not correct because it best approximates the costs that correspond to the extent to which an agent's unjustified and unexcused contributions individually increase expected harm.

The Share of the Total Account is not supported, then, by either of the two arguments discussed in this section. This does not mean, of course, that the account is false. Over the next two sections, I will argue that the Share of the Total Account is false.

[13] Note that the relationship between total greenhouse gas emissions and total actual harm is a nonlinear threshold function. Yet because the precise location of the relevant threshold-levels is unknown, the relationship between emissions and *expected* harm is roughly linear.

V. The Share of the Total Account and the Joint Liability Account

To better evaluate the Share of the Total Account, it will be helpful to introduce a competing account: what I will call the *Joint Liability Account*. The Joint Liability Account holds that (CJ-MA) allocates the total costs of rectifying a collectively-caused harm or threat among current duty-bearers in proportion to the extent of each agent's unjustified and unexcused contributions to the harm or threat. Therefore, the Joint Liability Account implies that in *Dead Polluter*, each B, C, and D has a duty to bear one-third of the $1,200 magnitude harm that results from the actions of A, B, C, and D (or $400 each). If B were to die, each C and D would have a duty to bear one-half of $1,200, or $600 each.

Let's consider the merits and deficiencies of the Joint Liability Account and the Share of the Total Account. Interestingly, Caney (2005) briefly considers but quickly rejects something like the Joint Liability Account, at least as it applies to dead emitters. He writes:

> The suggestion, then, is that current polluters should pay the costs of their pollution and that of previous generations. In this way, it might be said, the mitigation and adaptation costs of climate change are shouldered by the polluters (and not by non-polluters). But this seems unfair: they are paying more than their due [. . .] It is alien to the spirit of the [polluter pays] principle to make people pay for pollution which is not theirs. (2005, p. 760)

I understand Caney to be arguing that it would be unfair for the (currently living) duty-bearers identified by the polluter pays principle to bear the entire cost of preventing or compensating for climate change-induced harm, given that their emissions compose only a portion of the total emissions causing harmful climate change. If we generalize away from the circumstances of climate change, Caney's argument suggests that the Joint Liability Account unfairly burdens some agents by requiring them to bear all the costs of rectifying a harm or threat when these agents only contribute a portion of the total contributions that cause the harm or threat.

It is helpful, I believe, to point out that Caney's criticism of the Joint Liability Account appears to presuppose a particular interpretation of (CJ-MA)'s proportionality constraint. The proportionality constraint, recall, states that the costs that each agent incurs a duty to bear in service of rectifying a harm or threat must be proportionate to the extent of the agent's unjustified and unexcused contributions to the harm or threat. Caney suggests that if an agent bears a greater share of the costs than her share of contributions, as the Joint Liability Account prescribes she

must, then the costs that the agent bears is not proportionate to the extent of the agent's contribution to the harm or threat—it is to make them "pay for pollution that is not theirs." Caney's argument suggests, then, that the proportionality constraint should be interpreted in a way that implies the Share of the Total Account.

But we need not endorse Caney's interpretation of the proportionality constraint. We are ultimately concerned with the question of whether some cost is proportionate to the extent of the agent's unjustified and unexcused contributions to a harm or threat in the sense that it would be *fair* for the agent to bear that cost, given the extent of the agent's unjustified and unexcused contributions to the harm or threat. Therefore, we should investigate the ways in which it might be fair (or unfair) for an agent to bear some cost, particularly when the fairness (or unfairness) is explained by the extent of the agent's unjustified and unexcused contributions to a harm or theat.

With respect to collectively-caused harms and threats, there are two ways in which it might be fair (or unfair) for an agent to bear some cost. First, an agent's bearing some cost might be fair (or unfair) to the agent in comparison to the costs borne by the other duty-bearers. For example, in *Waste*, suppose that A bears half the costs of rectifying V's harm, while B, C, and D together bear the other half (such that each B, C, and D bears one-sixth of the total cost). It would be unfair for A to bear this cost, given that each B, C, and D contributes to the harm to the same extent as A (and given that there are no other morally relevant differences among their actions).

Notice, however, that both the Share of the Total Account and the Joint Liability Account imply that the costs that agents incur duties to bear are fair in this way. For instance, in *Dead Polluter*, both the Share of the Total Account and the Joint Liability Account imply that B, C, and D should each bear the *same* costs as one another, given that each contributes to the harm to the same extent (and given that there are no other morally relevant differences among their actions). Whichever account is correct, there is no difference in the fairness of the allocation of costs *among* the agents who incur duties to bear costs on the basis of (CJ-MA).[14][13]

Let's consider, then, the second way in which it might be fair (or unfair) for an agent to bear some cost. It might be unfair for an agent to bear a cost *independently* of the costs that other agents have duties to bear on the basis of (CJ-MA). Consider:

[14] Might it be unfair that, according to the Joint Liability Account, B, C, and D must bear the cost of rectifying the harm rather than *A*? Perhaps there is something unfair about A's having avoided paying her fair share of the costs—that through death, A somehow "escaped" justice. However, even if it is unfair that A does not pay her fair share before her death, notice that this unfairness does not go away were the costs previously assigned to A borne by someone *other* than B, C, or D. That is, these costs must be borne by someone, whether by V, by the other contributors to the harm, or by third parties on the basis of some other allocative principle. So this kind of unfairness (if it even exists) can't support an objection to the Joint Liability Account that doesn't also apply to the Share of the Total Account.

> *Mostly Dead Polluters*: Each of one million agents negligently discharges a drop of industrial waste into the creek. The one million drops together result in one million dollars of damage, though no one agent's contribution is necessary or sufficient for the harm. None of the agents has either an excuse or a justification for her action. All but one of the agents die before discharging their duties. Gert is the only living contributor.

Although Gert surely has a duty of corrective justice to bear a portion of the million-dollar cost, it seems unfair for Gert to bear the entire cost. It seems particularly unfair if Gert were not very rich and would take her the rest of her life to raise the million dollars. The unfairness is not explained by the costs that other liable agents bear, given that there are no other liable agents still alive. Rather, the million-dollar cost exceeds some upper limit on what it is fair for Gert to bear in service of rectifying the harm, independently of the costs that other liable agents bear. I will call the maximum cost that it is fair for a liable agent to bear independently of the costs that other liable agents bear the agent's *maximum cost threshold*.[15]

Mostly Dead Polluters gives us good reason to reject the Joint Liability Account. The Joint Liability Account implies that Gert has a duty to bear the entire million-dollar cost. I have said that it is implausible that Gert has a duty to bear this cost. The cost exceeds Gert's maximum cost threshold.

Notice that the reason the cost exceeds Gert's maximum cost threshold is that the large cost does not seem to be in proportion to the extent of Gert's minimal contribution to the harm. However, accepting the judgment that the million-dollar cost is not in proportion to the extent of Gert's contribution to the harm does not commit us to accepting the conception of the proportionality constraint that implies the Share of the Total Account. It may be that the extent of an agent's unjustified and unexcused contributions to a harm or threat is one of *several* factors that determine the agent's maximum cost threshold, and hence whether it would be fair for the agent to bear some cost. If so, then perhaps an agent's maximum cost threshold can be greater than the one set by the Share of the Total Account.

VI. Rejecting the Share of the Total Account

An agent's maximum cost threshold can indeed be greater than the one set by the Share of the Total Account. It can fair for an agent to bear a cost that is greater—

[15] For the notion of a maximum cost threshold, I credit Tadros (2013; 2017: Ch. 4), who uses a similar concept (what he calls a "maximum *harm* threshold") in his account of liability to punishment. According to Tadros, an agent's maximum harm threshold is the maximum amount of harm the agent is liable to suffer as punishment.

sometimes significantly greater—than the cost that the Share of the Total Account implies that would be fair for the agent to bear.

Before I argue for that claim, notice that even in circumstances in which only one agent causes a harm, the cost that the agent has a duty of corrective justice to bear can be limited by her maximum cost threshold. Suppose, for example, that while walking on a busy street, I carelessly bump into you, causing you to sprain your ankle. Suppose that for whatever reason, it would cost one million dollars to restore your ankle to full strength and to compensate you for your pain. Suppose also that I am not very rich, and that it would take me the rest of my life to work off the debt. Although I clearly owe you *something*, it seems unfair that I must bear the entire million-dollar cost. My behavior, though careless, does not make me liable to bear such a large cost.[16] This suggests, then, that an agent's maximum cost threshold is not determined solely by the extent of an agent's unjustified and unexcused contributions to a harm or threat. In the above case, the fact that my action is negligent, rather than malicious, seems to be relevant to my maximum cost threshold. If I had maliciously caused you to sprain ankle, then it seems that I would have a duty of corrective justice to a bear a much greater portion of the million-dollar cost, perhaps even all of it. This suggests that the degree of an agent's *culpability* for causing or contributing to a harm or threat is relevant to the agent's maximum cost threshold, and not just the extent of the agent's unjustified and unexcused contributions. I will return to this idea below.

Perhaps when a harm or threat is collectively-caused, the cost that it would be fair for an agent to bear is *limited* by the extent of the agent's unjustified and unexcused contributions in the sense described by the Share of the Total Account. Perhaps other factors (like an agent's diminished culpability) can only reduce the agent's maximum cost threshold below the agent's share of total costs.

It is highly doubtful that this is the case, however. An agent's maximum cost threshold can exceed the maximum cost that the Share of the Total Account implies that it would be fair for the agent to bear.

To see why, consider first the following case:

> *Field Clearing 1:* Anton wants to clear his yard of a large boulder. The most convenient way for him to do so is to push the boulder off a nearby cliff. Unfortunately, V is unwittingly sitting in his car just below the cliff. Anton knows this but doesn't care. He pushes the boulder off the cliff, and it lands on V's car, gravely injuring V.

[16] For a discussion of this kind of case, see Waldron 1995. See also Tadros (2013: 302) for a similar case.

Anton, we can agree, incurs a duty of corrective justice to compensate V. Moreover, it is difficult to imagine a monetary cost that would be unfair for Anton to bear, if Anton's bearing that cost were necessary to compensate V fully for his injury. So although the cost that Anton has a duty to bear may be limited by *necessity* (by the cost required to fully compensate V for the injury) or by Anton's capacity to bear the cost, any monetary cost that is needed to rectify the harm to V would not exceed Anton's maximum cost threshold.[17]

Here is a brief argument for this claim. Any cost needed to rectify V's harm that is not borne by Anton must be borne by someone else, either by V himself or by third parties (perhaps on the basis of a different principle). Anton could have avoided bearing any rectificatory costs by fulfilling his moral obligation to refrain from rolling the boulder off the cliff. Because Anton culpably failed to fulfill his moral obligation, Anton has only a weak objection to bearing even extremely significant costs if his bearing that cost were necessary to compensate V for the harm. On the other hand, both V and third parties have strong objections to bearing any portion of the cost, given that there was very little that they could have done to avoid bearing it. So it is fair for A to bear even very significant costs to compensate V, rather than for V or third parties to bear any portion of the cost themselves.[18]

Yet now consider:

> *Field Clearing 2:* Same as *Field Clearing 1*, but Beth knows that Anton is going to push the boulder over the cliff. Hoping to get revenge on V, Beth coaxes V to park his car just below the cliff. Anton knows that Beth does this, but Anton pushes the boulder anyway (solely to remove the boulder from his yard). The boulder lands on V's car, gravely injuring V.

It seems that Anton's maximum cost threshold in *Field Clearing 2* is the same as in *Field Clearing 1*. In *Field Clearing 2,* Anton performs the same action, which results in the same foreseeable consequences as in *Field Clearing 1*. The fact that Beth contributes to the harm V suffers in *Field Clearing 2* does not seem to reduce the maximum cost that it is fair for Anton to bear. Again, perhaps the cost that Anton has a duty to bear is limited by Anton's capacity to bear it. It may also be limited by necessity, especially since the cost that Anton must bear to ensure that V is fully

[17] Perhaps we could imagine a non-monetary cost that exceeds Anton's maximum cost threshold. Suppose, for example, that the only way to restore fully V's abilities would be to kill Anton and harvest his organs. It is plausible that this cost would exceed Anton's maximum cost threshold. If that example doesn't convince you, imagine that for whatever reason Anton must first be tortured.

[18] For a discussion of the relevance of culpability to liability in the manner sketched here, see Scanlon (1998: Ch. 6) and Tadros (2011).

compensated may be only a portion—perhaps half—of the total cost needed to fully compensate V, given the existence of a second liable agent in *Field Clearing 2*. However, the cost does not seem to be limited by Anton's maximum cost threshold.

The conclusion reached here is already strong evidence against the Share of the Total Account. Suppose that in *Field Clearing 2*, (CJ-MA) implies that Anton and Beth each incur a duty to bear one-half of the cost needed to fully compensate V. Suppose, however, that Beth dies before she can contribute her share. Given that the cost of fully compensating V does not exceed Anton's maximum cost threshold, it seems that Anton should bear the entire cost himself.

The argument for this conclusion is similar to my previous argument for why it is fair for Anton to bear the entire rectificatory cost in *Field Clearing 1*. In *Field Clearing 2*, when Beth dies, Beth's share of the cost must be borne either by Anton, by V, or by third parties. Anton has only a weak objection to bearing the additional cost, given that he could have easily avoided bearing this cost by fulfilling his moral obligation to refrain from rolling the boulder off the cliff. Yet V and third parties have strong objections to bearing the cost; they could not reasonably have done anything to avoid bearing it. So it is fair for Anton to bear the entire cost of fully compensating V (including Beth's share), rather than for V or any third party bear any portion of the cost themselves.

Perhaps a proponent of the Share of the Total Account could respond by pointing out that in *Field Clearing 2*, both Anton and Beth are causally necessary for V's injury, whereas in the case of climate change, it is unlikely that any individual's action is causally necessary for any climate change-induced harm. Perhaps, then, the reason that the maximum cost threshold is the same in *Field Clearing 1* and *Field Clearing 2* is that, in both cases, Anton's action is causally necessary for V's harm, but in cases in which no agent's action is causally necessary for the harm, an agent's maximum cost threshold is limited by the extent of the agent's unjustified and unexcused contributions, in the sense implied by the Share of the Total Account.

We can investigate this possibility. Consider first:

> *Field Clearing 3*: Both Anton and Beth want to clear their respective yards of a medium-sized boulder. The quickest way for each to do so is to push his or her boulder off the same nearby cliff. Unfortunately, V is unwittingly sitting in his car just below the cliff. Anton knows that if he (and he alone) pushes his boulder off the cliff, it will not harm V. However, he knows that if both he and Beth push their respective boulders off the cliff, the combined weight of their boulders will land on V's car, gravely injuring V. Anton knows that Beth will push her boulder. Both Anton and Beth push their boulders, which land on V's car, gravely injuring V.

> Anton and Beth do not intentionally coordinate their actions, and each is motivated solely to clear his or her boulder.

The only difference between *Field Clearing 2* and *Field Clearing 3* is how each Anton and Beth make a causally necessary contribution to the harm that V suffers (Beth by pushing a boulder, Anton by pushing a medium-sized boulder rather than a large one). It is hard to see why this difference should matter to Anton's maximum cost threshold. If Anton's maximum cost threshold is the same in *Field Clearing* 3 as in *Field Clearing* 2, then by transitivity, Anton's maximum cost threshold is the same in *Field Clearing 3* as in *Field Clearing 1*.

Now consider:

> *Field Clearing 4:* Same as in *Field Clearing 3*, but there are *three* agents—Anton, Beth, and Clarice—who wish to clear their respective yards of a medium-sized boulder. If any two of the three agents push their boulders off the cliff, V will (foreseeably) be gravely injured. Anton knows for certain that Beth and Clarice will push their boulders. All three agents push their boulders, which land on V's car, gravely injuring V. None of the agents coordinate their actions, and each is motivated solely to clear their respective boulders, such that each would have pushed his or her boulder had the others refrained from doing so.

Notice that the only difference between *Field Clearing 3* and *Field Clearing 4* is the addition of Clarice's contribution. Yet it is hard to see why the mere addition of Clarice's contribution should make a difference to Anton's maximum cost threshold, especially given that Anton would not have changed his behavior even had his action been causally necessary for the harm. Yet if I am correct that the maximum cost threshold for Anton is the same in *Field Clearing 4* as it is *Field Clearing 3*, then again by transitivity, Anton's maximum cost threshold in *Field Clearing 4* is the same as in *Field Clearing 1*. This is so even though in *Field Clearing 4*, A's contribution to the harm to V is not causally necessary for the harm.

In *Field Clearing 4*, (CJ-MA) implies that Anton, Beth, and Clarice should split evenly the costs of fully compensating V. (assuming that each is able to do so). Yet, suppose that Beth and Clarice die before contributing their share of the costs. Because Anton's maximum cost threshold is the same in *Field Clearing 4* as in *Field Clearing 1*, then it is fair for Anton to bear the entire cost of rectifying the harm to V, including Beth and Clarice's shares. Once again, Anton's objection to bearing this cost is weak, given that he could have easily avoided bearing the cost simply by refraining from pushing his boulder. On the other hand, V and third parties could

not have reasonably avoided bearing the cost. Because it is fair for Anton to bear the entire cost, rather than V or any third party to bear even a portion of the cost, Anton has a duty of corrective justice to bear the entire cost.

My argument does not imply that the extent of an agent's contribution to a harm is never relevant to the agent's maximum cost threshold. We saw in *Mostly Dead Polluters* that the extent of Gert's contribution to a harm is relevant to her maximum cost threshold. Yet as *Field Clearing 4* suggests, the extent of an agent's contribution does not limit the agent's maximum cost threshold in the sense implied by the Share of the Total Account. It is plausible, then, that in *Mostly Dead Polluters*, although Gert does not have a duty to bear the entire million-dollar cost of rectifying the harm she contributed to causing, her maximum cost threshold is significantly greater than cost she would have a duty to bear according to the Share of the Total Account. Consider again that whatever cost Gert does not bear will have to be borne by someone else—someone who is not responsible for helping create the harm. Accordingly, it seems fair that Gert should bear a significant cost to rectify the harm—significantly more, at least, than the $1 prescribed by the Share of the Total Account—even if it would be unfair for her to bear the entire cost herself.

The Share of the Total Account is false because it implausibly limits an agent's maximum cost threshold to the cost that is strictly proportional to the extent of the agent's contributions to a harm or threat. I have also argued that the Joint Liability Account is false because it implies that there is no maximum cost threshold. Yet we can easily modify the Joint Liability Account in light of my discussion above. According to what I will call the *Modified Joint Liability Account,* (CJ-MA) allocates the total costs of rectifying a collectively-caused harm or threat among all current duty-bearers; however, the cost that any agent has a duty to bear cannot exceed the agent's maximum cost threshold. An agent's maximum cost threshold is determined partly by the extent of the agent's unjustified and unexcused contributions to the harm or threat, but other factors, like an agent's culpability, are relevant as well.[19]

VII. Evaluating the Insufficiency Claim

Let's now return to the Insufficiency Claim. In the previous section, I argued that the Modified Joint Liability Account is correct: (CJ-MA) allocates the total costs of rectifying a collectively-caused harm or threat among all current duty-bearers, but

[19] Note that my arguments for the Modified Joint Liability Account have implications for the fairness of the distribution of costs *among* duty-bearers. If factors other than the extent of an agent's unjustified and unexcused contributions are relevant to determining the agent's maximum cost threshold, these other factors would also seem to be relevant to the cost that it is fair for an agent to bear in comparison to other duty-bearers.

the cost that any agent has a duty to bear cannot exceed the agent's maximum cost threshold. Because an agent's maximum cost threshold can be greater than an agent's capacity to bear that cost, the cost that (CJ-MA) assigns an agent a duty to bear is also limited by the agent's capacity to bear that cost.[20] Therefore, for a collectively-caused harm or threat, if the sum of every current duty-bearer's (capacity-limited) maximum cost threshold is less than the total costs of rectifying the harm or threat, then (CJ-MA) does not assign duties to bear all of the rectificatory costs.

With respect to the circumstances of climate change, then, if the sum of every current duty-bearer's (capacity-limited) maximum cost threshold is less than the total cost required to prevent or compensate for all climate change-induced harm, then the polluter pays principle does not assign duties to bear all of these costs, i.e., the Insufficiency Claim is true.

How, then, should we evaluate whether the sum of every current duty-bearer's (capacity-limited) maximum cost threshold is less than the total costs needed to prevent or compensate for all climate change-induced harm? One method is to estimate the *average* cost that current duty-bearers would have a duty to bear were the polluter pays principle to assign duties to bear the total costs needed to prevent or compensate for all climate change-induced harm. We can then judge whether it would be fair for an average current duty-bearer to bear this cost, given factors like the magnitude of the climate threat, the extent an average duty-bearer's unjustified and unexcused emissions, and the degree of culpability an average duty-bearer has for producing these emissions. If it seems like the average cost would be fair for an average current duty-bearer to bear, then this is good evidence that the Insufficiency Claim is false.

To employ this method, we first we need to estimate the total costs needed to prevent or compensate for all climate change-induced harm. Any estimate of these costs undoubtedly makes several controversial empirical and normative assumptions. I avoid addressing these assumptions directly by simply appealing to estimates provided by others. There are two kinds of preventative costs for which there are reasonably reliable estimates: mitigation costs (the costs of preventing further changes to the global climate system) and adaptation costs (the costs of protecting persons against potentially harmful climate change impacts). I shall ignore compensation costs, which at this time are comparatively small (but will become larger as climate change worsens).

[20] I assume that an agent's capacity to bear a cost can be limited both by physical and moral considerations. Regarding the latter, if bearing a cost would drop an agent below a certain level of well-being, it may be wrong to require the agent to bear the cost, even if the cost is below the agent's maximum cost threshold.

According to a 2014 report published by the International Energy Agency, the costs of preventing increases in global temperatures from exceeding two degrees Celsius by decarbonizing the global economy would be approximately 44 trillion dollars through 2050 (IEA, 2014).[21] In addition, the United Nations Environmental Programme estimates that developing countries will require approximately 500 billion dollars per year in adaptation financing by 2050 to protect against the worst impacts from climate change (UNEP, 2016).[22] So, that's nearly 61 trillion dollars over the next 33 years or so. Let's round these costs up to 99 trillion dollars to account for the large degree of uncertainty in these estimates (and to make the math easy). So 99 trillion dollars over 33 years is approximately three trillion dollars per year until 2050. For context, global GDP in 2016 alone is approximately 75 trillion dollars (World Bank, 2016).

Suppose that around one billion currently living individuals both (a) have made significant unjustified and unexcused contributions to global climate change, and (b) have the capacity to bear reasonably significant rectificatory costs. There are approximately one billion people living in all OECD Annex I countries.[23] If the polluter pays principle were to assign duties to bear the entire 99 trillion dollar cost among these one billion persons, the average cost that each person would have a duty to bear is 99,000 dollars over the next 33 years, or about 3,000 dollars per year per person until 2050.

While 3,000 dollars per person per year is not insignificant, I do not think that it would be unfair for an average person living in a wealthy country to bear this cost, such that the cost exceeds the person's maximum cost threshold. Consider, first, that unless prompt and significant preventative action is taken, climate change threatens to cause the deaths and suffering of hundreds of millions of people over the next century (IPCC, 2014).[24] The extent of the harm that is threatened by the unjustified and unexcused contributions of the one billion liable agents is therefore extremely large, and is probably the greatest current foreseeable threat to fundamental human interests.

---

[21] Importantly, this estimate does not include the vast economic benefits that will be gained by avoiding dangerous climate change. Although most of these benefits will accrue to future generations, some of them will accrue to members of the current generation.

[22] There is likely some overlap in these estimates. To wit, the more we mitigate climate change, the less adaptation we will need (and vice-versa). The estimate of the adaptation costs assumes that we will exceed the two degree threshold.

[23] Of course, many of the one billion persons living in Annex I countries are children. However, there are also many persons living in non-Annex I countries who are reasonably wealthy and who have made unjustified and unexcused causal contributions to harmful climate change. I will assume the two groups cancel each other out.

[24] This may underestimate the extent to which people will suffer and die from climate change. See, e.g., Nolt (2011).

Second, even if the one billion liable agents' emissions constitute only a portion of the total emissions that have resulted in harmful climate change, this portion is not insignificant. Consider that nearly half of all greenhouse gas emissions generated between 1850 and 2012 were generated between 1990 and 2012—that is, *after* agents knew, or ought to have known, that greenhouse gas emissions were causing harmful climate change (WRI, 2014). Moreover, approximately half of these post-1990 emissions come from the United States, the Russian Federation, European Union countries, Canada, and Japan, whose citizens compose the majority of the one billion liable agents (WRI, 2014). So about twenty-five percent of cumulative greenhouse gas emissions were generated by citizens of these countries, after 1990.

Third, given how wealthy many of the citizens of these countries are (and have been since at least 1990), many of the greenhouse gas emissions produced by these citizens after 1990 were likely neither justified nor excused. This means, then, that many of the citizens of these countries had excellent opportunities to reduce their post-1990 emissions, but chose not to do so. These individuals, it seems to me, are highly culpable. It seems fair, then, that many of these individuals should bear considerable costs to rectify the climate change-induced harms and threats that they have made significant unjustified and unexcused contributions to causing.

Fourth, had wealthy countries and their citizens acted to mitigate climate change when they first learned that climate change was occurring, the costs of mitigating and adapting to climate change would have been *considerably* less than current costs. For example, in just two years (between 2012 and 2014), the International Energy Agency estimates that the cost of decarbonizing the global economy (to keep climate change under two degrees Celsius) increased by eight trillion dollars, or 25% over that time (IEA, 2014). Surely the maximum cost thresholds of the one billion current duty-bearers must reflect the fact that, collectively, they had an opportunity to bear much smaller costs to rectify harmful climate change but chose not to do so.

Given these considerations, it seems to me that, on average, 3,000 dollars per year per person does not even come close to exceeding the cost that it is fair for the average current duty-bearer to bear in service of rectifying harmful climate change. In other words, it does not exceed the average current duty-bearer's maximum cost threshold. It is likely, then, that the Insufficiency Claim is false.

IX. Conclusion

In this paper, I argued that the Insufficiency Claim is likely false. The total costs that the polluter pays principle assigns agents duties to bear are likely sufficient to cover

the total costs that are required to prevent or compensate for all climate change-induced harm.

In closing, it is important to note that even though the Insufficiency Claim is false, this does not mean that the only duties that agents have are those assigned by the polluter pays principle. In addition to having duties of corrective justice, it is plausible that many agents also have duties of distributive justice and duties of beneficence to bear costs. But the costs that agents have these duties to bear are not necessarily separate from, or in addition to, the costs that agents have duties of corrective justice to bear. Rather these duties overlap, giving agents powerful, all-things-considered moral obligations to bear significant costs to address climate change.

References

Baatz, C. (2013). Responsibility for the past? Some thoughts on compensating those vulnerable to climate change in developing countries. *Ethics, Policy & Environment,* 16 (1), 2013, 94–110.

Bell, D. (2010). Justice and the politics of climate change. In C. Lever-Tracy (Ed.), *Routledge Handbook of Climate Change and Society* (pp. 424–441). New York, NY: Routledge.

Broome, J. (2012). *Climate matters: Ethics in a warming world*. New York, NY: W.W. Norton.

Caney, S. (2005). Cosmopolitan justice, responsibility, and global climate change. *Leiden Journal of International Law*, 18(4), 747–775.

Caney, S. (2010). Climate change and the duties of the advantaged. *Critical Review of International Social and Political Philosophy,* 13(1), 203–28.

Cripps, E. (2013). *Climate change and the moral agent: Individual duties in an interdependent world*. Oxford: Oxford University Press.

Duus-Otterström, G. (2014). The problem of past emissions and intergenerational debts. *Critical Review of International Social and Political Philosophy*, 17(4), 448–469.

Hiller, A. (2011). Climate change and individual responsibility. *The Monist*, 94(3), 349–368.

IEA. (2014). *Energy technology perspectives: Executive summary.* International Energy Agency, Paris: IEA Publications.

Morgan-Knapp, C., & Goodman, C. (2015). Consequentialism, climate harm and individual obligations. *Ethical Theory and Moral Practice*, 18, 177–190.

Meyer, L. & Roser, D. (2010) Climate justice and historical emissions. *Critical Review of International Social and Political Philosophy*, 13(1), 229–253.

Moellendorf, D. (2014). *The moral challenge of dangerous climate change: Values, poverty, and policy.* Cambridge: Cambridge University Press.

Page, E.A. (2008). Distributing the burdens of climate change. *Environmental Politics,* 17(4), 556–575.

Parfit, D. (1984). *Reasons and persons*. Oxford: Clarendon Press.

Scanlon, T.M. (1998). *What we owe to each other*. Cambridge: Harvard University Press.

Shue, H., 1999. Global environment and international inequality. *International Affairs,* 75 (3), 533–537.

Tadros, V. (2011). *The ends of harm*. Oxford: Oxford University Press.

Tadros, V. (2013). Responses. Law and Philosophy 32, 241–325

Tadros, V. (2017). *Wrongs and Crimes*. Oxford: Oxford University Press.

UNEP. (2016). *The adaptation gap: Finance report.* United Nations Environment Programme (UNEP), Nairobi, Kenya.

Vanderheiden, S. (2008). *Atmospheric justice: A political theory of climate change.* Oxford: Oxford University Press.

Waldron, J. (1995). Moments of carelessness and massive loss, in D. Owen (Ed.), *Philosophical Foundations of Tort Law* (pp. 387–408) Oxford: Clarendon Press.

WRI. (2014). *Climate analysis indicators tool: WRI's climate data explorer*. World Resources Institute. Retrieved from http://cait2.wri.org.

WHO. (2017). Climate change and health. World Health Organization. Retrieved from http://www.who.int/mediacentre/factsheets/fs266/en/

Zelletin, A. (2014). Compensation for historical emissions and excusable ignorance. *Journal of Applied Philosophy* 32(3), 258–274.

Martin Kolk[1]

Demographic Theory and Population Ethics – Relationships between Population Size and Population Growth

Demographic theory aims at explaining how population systems regulate themselves given available resources. Population ethics is concerned with demography in the sense that the analytical objects of interest are births, deaths, and populations. However, demographic theory which explores theoretically when, how and why populations grow, based on empirically observed patterns, has up until now played a minor role in population ethics. Similarly, debates about population dynamics among demographers have seldom been concerned with ideas and concepts in population ethics. In this manuscript, I will give a brief outline of how population size, population growth, and welfare mutually affect each other. Theories on the endogeneity between population size, population growth, and welfare will be referred to as demographic theory. I will give a particular focus on how population growth responds with respect to welfare, as welfare, utility, well-being, and happiness are important concepts in population ethics. A key concept in demographic theory is population homeostasis (the dynamics of a system which maintains a population at a steady population size, or growth rate), in particular resource dependent homeostasis. I will also discuss demographic theory in relation to historical and future demographic change.

[1] Demography Unit, Stockholm University, Centre for Cultural Evolution, Stockholm University, and the Institute for Futures Studies, Stockholm, martin.kolk@sociology.su.se. Financial support by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences) is gratefully acknowledged.

Introduction

Population ethics is clearly concerned with demography in the sense that the analytical objects of interest are births, deaths, and populations.[2] However, demographic theory which explores theoretically when, how and why populations grow, based on empirically observed patterns, has up until now played a minor role in population ethics. Similarly, debates about population dynamics among demographers have seldom been concerned with ideas and concepts in population ethics. In this manuscript, I will give a brief outline of how population size, population growth, welfare, mutually affect each other. Theories on the endogeneity (co-dependencies) between population size, population growth, and welfare will be referred to as demographic theory. I will discuss how population growth responds with respect to welfare, as welfare or utility is an important concept in population ethics. I will discuss population homeostasis (the dynamics of a system which maintains a population at a steady population size, or growth rate), with a focus on resource dependent homeostatic mechanisms. I will pay particular attention to negative relationships between population size and welfare, and refer to such relationships as "Malthusian", after Thomas Malthus.

Social scientists have over centuries discovered regularities in how populations grow and shrink, and developed methods to measure and model these accurately. Compared to many other domains of human interaction demographic theory is concerned with empirical phenomena such as births and deaths that are easy to measure and define, also across widely different contexts and times. Demographic theory is concerned with describing populations as analytical dynamic systems, and to find empirical patterns in how and why populations change. As such, it is descriptive and analytical and is not focused on if outcomes are preferable or more desirable than an alternative outcome. While economic demography is concerned with welfare or available resources, it is often related to resources in an ecological sense, where resources regulate population growth and size, and where the dynamics and outcomes are unrelated to ethical outcomes. There are strong overlaps between population studies and evolutionary biology, which both model births, deaths, population growth, and the relative size of different groups over time. In fact, biological evolution can be viewed as a demographic process where births and deaths change the distribution of traits in each new generation (population growth is a central concept both in ecology and evolutionary biology). Theorizing on

[2] Throughout this text I will use the term population and population size to refer to the number of currently living individuals at a single time point. The geographical area or definition will be left unspecified, and I will occasionally use population density analogously to population size, as for a certain (closed) area - such as Earth - the two concepts are identical.

evolution and human population systems have a shared scientific ancestry before the 20th century.

Models in demography theory assume that individuals are agents whose actions are both affected by and affect well-being, but that motivations are related to individual choices and often don't maximize population wellbeing. Therefore, the long-term population outcome of demographic systems typically is not what maximizes either average or aggregated welfare in a population. Population equilibriums in demographic theory are therefore different from calculations of optimal population size, which aims at finding the population size that maximizes welfare (from either an average individual's or a cumulative welfare perspective). As demographic change is driven by choices made dependent on individual (average) welfare, population systems will be determined by the average welfare of an individual and not by global aggregated utilitarian welfare. In this manuscript I will mostly discuss welfare from an average perspective and not from the total utilitarian perspective in population ethics (cf. Singer, 1976), however it is easy to move from an average welfare to a total welfare perspective as it simply consists of multiplying average welfare with population size (though this has interesting implications for population models where population size can change over time).

A theoretical understanding of how populations grow, decline and maintain equilibriums over long time spans are relevant for many topics in population ethics. Many of the assumptions of Population ethics, such as Malthusian assumptions of average welfare declining with additional individuals are implicit, and ethicists may want to explore other assumptions. A nuanced understanding of demographic theory may also help to understand if possible imagined scenarios are unlikely, or even nearly impossible empirically. If a theory produces possible paradoxes and unacceptable conclusions that are implausible based on demographic theory, they may be less of a concern for a theory, than other scenarios which are more grounded in demographic theory. Demographic theory may also inform population ethicists if certain population scenarios are plausible only in the short term, or if they are plausible over longer time scales.

While population ethicists can and should make thought experiments on any imaginable population and welfare combinations, understanding the basic parameters in which populations, population growth, and welfare interacts helps to ground theorizing in population ethics and makes it more applicable to empirical circumstances. There is a long (sometimes non-explicit) tradition in population ethics to assume negative welfare to larger populations following a Malthusian perspective.[3] Broadly, one can distinguish two different approaches in population

[3] See for example the following three quotes from population ethicists. In the quote below by Parfit (1984) he, a) takes a Malthusian relationship as an empirical fact in some contemporaneous societies,

ethics when reasoning about the effect of population size and welfare. Good illustrations of the first approach are Parfit (1984), Arrhenius (2011), and Broome (1996). They compare hypothetical populations to each other, and often the analytical treatment allows both a positive relationship between population size and welfare, as well as a negative relationship. However, typically the framing and examples of their models are to assume an explicit negative function between population size and welfare. An example of this is Parfit (1984). A reason for the implicit Malthusian settings in this school of population ethics, is that the opposite case, of increasing welfare in larger populations, does not involve trade-offs with larger populations and value infinite population sizes. A second approach can be illustrated with Dasgupta (1998) where authors create explicit models to calculate welfare based on population size. Such models typically make and motivate Malthusian relationships between population size. This illustrates how having a thorough understanding of demographic theory, and both Malthusian and non-Malthusian perspectives in population ethics, may clarify implicit assumptions in population ethics and may open up new avenues of research.

The main value of demographic theory for population ethics is through the understanding that population growth (and with some delay, population size) will respond to current welfare in a population. For example, a hypothetical very large population with highly negative welfare is likely to over time revert to lower population sizes with higher welfare. Similar over long periods of time, both very small and very large populations are likely to move towards intermediate popu-

---

and b) then after first qualifying that a Malthusian relationship might not apply everywhere (in two sentences) c) proceeds discuss the Malthusian case and implications of such a relationship (over 59 pages).

> "The effect of population growth on existing people – My couple assume that the existence of an extra child would not on balance be worse for other people. In many countries, in many periods, this has been true. But in many other periods it has not been true. In these periods, if there had been more people, people would have been worse off. This is now true in many countries. In these countries, if the population grows, the quality of life will be lower than it would be if the population did not grow. These are the cases that I shall discuss" Parfit (1984, p. 382).

The quote below is the basis of a thought experiment introducing optimal population from Arrhenius and Campbell (2017):

> "If the current generation continues to consume resources at the expense of future generations, and the population increases significantly, this could lead to an enormous population—ten billion people per generation—in which most people's lives are barely worth living. Suppose we could instead create a smaller population—around one billion people per generation—with very good lives. Which population would be better?"

This is the representation of the world that Dasgupta (1998) uses to discuss the value of additional people in an ecological model on optimal population size:

> "More importantly is the thought that, like everyone else, potential geniuses, if they are to flower, require resources, such as potable water, food, clothing, medical care and education. So it would seem that we cannot get very far with the idea of a desirable population size without considering resource constraints. Whether it be for a household or for an entire nation, population policy cannot be formulated without a concurrent saving and investment policy."

lation sizes. Such insights are most important when population ethicists are interested in cumulative welfare across a large number of generations. Understanding such relationships are valuable for population ethicists, as inquiries into optimal population size for human welfare will have to make assumptions on how current population size is related to welfare, how welfare is related to population growth, and how current population size is related to future population growth. Demographic theory might also be a useful perspective to examine trade-offs between reproductive rights and optimal population size, as a population with high fertility due and unrestrained reproductive rights in some cases may be described well by resource dependent demographic theory.

A fundamental insight from demographic theory is that all population systems over many generations are going to have homeostatic characteristics. In a homeostatic system, population increases and declines will eventually revert to an equilibrium, level. Long-term unbounded population growth is unlikely; as exponential growth quickly results in an unsustainable and implausible population size. Similarly, a homeostatic population model implies that population growth rate is positive at low population sizes, and that extinctions of populations thus are unlikely.

In a Malthusian model, homeostatic regulation is related to the availability of resources. When resources are plentiful the population will grow, and when they are scarce the population will decrease. This forms a basic framework for most theorizing on both human and animal populations. Economic demographers have further theorized on positive relationships between population size and population growth, under some circumstances. Such positive effects would stem from dynamic gains from for example specialization of labor, and increases in communication and innovation in dense populations. Given that such positive feedback loops are reasonable for some population sizes, it is possible to develop models combining Malthusian negative feedbacks and positive density-dependent feedbacks applicable at different population sizes. Implications of such theories for population ethics, as well as an empirical description on the determinants of contemporary and historical population systems will be discussed in the current manuscript.

The usage of terms such as welfare, utility, wages, and resources differs between subjects. In this text, I will use all of them largely interchangeably depending on which context appear suitable, in all cases to refer to access to material resources in a broad sense.[4] When I refer to the welfare of populations, I refer to the average welfare of an individual in that population (and not the aggregated welfare of that population). Because as I assume homogenous populations this will be the same as

[4] Note, that this means that it is clearly different (though likely to some degree overlapping) from a concept of welfare that also include non-material sources of happiness.

the welfare of any individual. Throughout this overview, the time perspective will be comparatively large and encompass multiple generations of humans (when I use the term long-term I refer to at least 5+ generations, longer perspectives than what is used in most governmental and scientific projections on population size). The determinants of population growth, mortality, and fertility in a shorter time perspective are different and are not the focus of this overview.

In this manuscript, I will first discuss the role of exponential growth, and how it relates economic demographic theory and population homeostasis. This will be followed by a presentation of some simple models in demographic theory, and after that examine implications for population ethics of such relationships. Finally, I will discuss the determinants of population dynamics of past, present, and future human populations.

Exponential growth, economic demography and population homeostasis

A fundamental characteristic of (human) populations is that populations grow (and decline) exponentially. This creates the potential for very quick growth, and decline when analysed over multiple generations. All human (and other animal) populations will, given favourable circumstance, quickly expand at such a quick rate that there after a longer time perspective must be some other factor holding back growth. For example, the French-speaking population in Quebec had population growth rates from natural [5] growth of around 2% for several centuries which increased the population several hundred times starting from the 17th century (Charbonneau, Desjardins, Légaré, & Denis, 2000).

That population size eventually stabilizes after a period of growth is a phenomenon that is repeatedly observed both in animal populations and for historic human population. This realization goes back to Malthus and served as inspiration for what Darwin called the "struggle for existence". Negative exponential growth is similarly powerful and can decrease population size very quickly. Even in a population where no individual ever dies would quickly stop increasing if each woman gave birth to less than two children during her lifetime, [6] as each new generation would be smaller than the former. For example, a population in which all

[5] Natural growth refers to population changes arising from births and deaths, not taking migration into account.

[6] Demographers us the concept of Net reproductive rate (NRR), which is typically defined from only a female perspective, and is the average number of surviving daughters of a women, at a given a level of mortality and fertility of the woman. If the NRR is larger than 1 a population will grow, and if it is smaller than 1 it will shrink. For example, a NRR of 1 would arise if a women on average give birth to two children, a sex ratio of 0.5, and that no woman dies before the end childbearing ages.

women have one child (or half of all women have a daughter) in their life would stabilize at twice the original size in a population where no one ever dies.[7]

Any fast-growing human populations will stop growing rapidly at some point due to the nature of exponential growth. The world population growth rate observed in the 1960s of 2.2% implies that earth's population would be 53,143 times bigger if this growth would be constant for another 500 years.[8] Most population can grow at times at similar speeds, but such growth rates will inevitably come to an end. A useful interpretation of the demographic transition, the shift from a high fertility-high mortality society, to a society with reduced mortality and consequently high population growth, to a low fertility-low mortality society, can be seen as an inescapable outcome of a process in which exogenous factors reduce mortality permanently. The demographic transition is then a (homeostatic) way for a society to once again reach population balance and escape unsustainable population growth (Wilson & Airey, 1999). A very demographic definition of overpopulation can be said to be the population size for which population growth turns negative.

In a homeostatic population model, population growth must in some way be related to population size itself. That populations are bounded by available resources is a key insight of Malthus in the 18th century,[9] and that human population growth reacts favourably to abundant resources, while decreasing when resources are scarce. A Malthusian relationship is central in ecological models applicable to all species, and there is clear evidence that such mechanisms regulate population size in historical agricultural and hunter-gatherer societies (Lee, 1987). This is an example of population homeostasis, in which population growth responds to population size,[10] in this case through a positive association between resource availability and population growth and a negative association between resource availability and population size. As discussed earlier, Malthusian assumptions remain dominant when population ethicists have incorporated demographic theory in demographic models.

---

[7] This results from a NRR of 0.5, as

$$a + ar + ar^2 + ar^3 + ar^4 + \cdots = \sum_{k=0}^{\infty} ar^k = \frac{a}{1-r}, \text{ for } |r| < 1.$$

gives a sum of 2 if a is equal to 1 and r equal to 0.5. This geometric series represents cumulative population when every new generation is half the size of the previous one.

[8] After 2000 years there would only be enough matter on Earth for 1 gram per person.

[9] Though ideas of a positive relationship between number of people and wealth have been common in many societies long before Malthus, and such ideas where a part of a broader scientific discourse in 18th century Europe (Hutchinson, 1967).

[10] A stable population equilibrium is when all individuals are living at subsistence level, but it is possible that the function between welfare and population growth is different, and that for example societies would regulate reproduction so that a population equilibrium would take place at welfare levels above subsistence minimum.

However, a Malthusian model cannot explain how human population size has increased from millions to billions in the last few thousand years. In the last 10,000 years the global population has not only grown exponentially, but in addition, the growth rate itself has been growing exponentially (Livi-Bacci, 2007). This implies that there must have been some development of human culture that has allowed an increasingly large population size over time, breaking a strict population homeostasis. Population growth seems to be in some way be related to population size itself in some way. It is plausible that an increasing population size or density[11] can also be associated with increasing living standards and population growth. A higher population density encourages specialization, which can be an efficient way to organize labour, encourage higher human capital. Higher population density may also increase the demand for new technology.[12] Transportation, communication, and access to ideas would also increase with increasing population density. Similarly, if innovations appear largely by chance, more people imply more novel ideas and innovations. A larger population can also maintain more knowledge at the same time (S. Ghirlanda, Enquist, & Perc, 2010; Lehmann, Aoki, & Feldman, 2011). Throughout the 19th century, a number of political economists criticized Malthus[13] and suggested that increasing population pressure might stimulate adoption and innovation that could counter diminishing returns (Hutchinson, 1967, pp. 152-202). The economist Ester Boserup examined the relationship between population density and innovation in Sub-Saharan Africa (Boserup, 1965, 1981), and found that higher population density on average increased the adoption of better agricultural techniques and technology, and increased both welfare and population growth. Any such positive dependencies between population size and welfare will be referred to as Boserupian in contrast to a negative dependency in a Malthusian model.

To illustrate both Boserupian and Malthusian relationships, consider as an example the population of medieval Europe. The population varied around 50 million inhabitants between 1000 and 1500 C.E (Livi-Bacci, 1999). If we assume a population over 700 million, similar to the population in the 2010s, living standards

[11] Note that throughout this manuscript I will treat population density and population size as analogous concepts. AIn the short-term populations will likely resort to migration when facing resource constraints, but in the longer run the destination population will face similar demographic pressures. At equilibrium population density and population size can then be treated as analogously. Most obviously Earth itself is for the foreseeable future a closed population system, in which population density and population size are identical. However, migration remains an important practical concern when attempting to test population theories using empirical data.

[12] When using technology throughout the article it is defined in very broad terms as any cultural institution, idea, capital investment, or way of live, which is associated with productivity. As such, it also reflects institutional and cultural change broadly and not only technical engineering inventions.

[13] Thomas Sadler in 1830, Friedrich List in 1841, and John McChulloch in 1846, all clearly and forcefully argued that better innovation and better agricultural techniques could and would be applied to counter diminishing returns (Hutchinson, 1967).

would have been very low. If, however, the population would only have been 10 000 individuals in all of Europe, it seems clear that such a population would not have been able to maintain the specialization and technology that Europe had during the middle ages, and that all the spare land and abundant resources would have been of relatively little use. A higher living standard, and plausibly higher population growth, would lie somewhere in between those extreme cases. Accounts of the economic and demographic history of Europe in the middle ages makes it plausible that Europe was close to its maximum population size given available resources around 1340, before the plague reduced the population of Europe (e.g. Livi-Bacci, 1999). The large population size in European before the plague resulted in that much sub-standard land and over labour-intensive cultivation was used, and that average living standards would have been higher at a lower population density. This implies that at some point, between 50 million and 10 000 average living standards would have been higher than at 50 million, though exactly at what point is very hard to estimate. It also implies that at the lower bound of that range, an increasing population, increased welfare and population growth, while at the upper bound (and above that range) the opposite would be true.

A homeostatic Malthusian model in which population growth is negatively correlated population size, and growth is high at low population size, and negative at higher population sizes, is described in figures 1-3. A key assumption of a Malthusian model is that higher population size is associated with lower welfare (figure 1). The relationship between welfare and population growth is assumed to be positive as vital rates (births and deaths) respond to greater abundance of resources (figure 2). This gives a negative relationship between population size and population growth (figure 3). The point at which population growth is about 0 is the equilibrium population size, around which a population would oscillate. In an animal population this would be the carrying capacity of the population, and this is also plausible for humans. However, both industrial and pre-industrial human populations may regulate reproduction in ways that would not necessarily result in a population on the brink of population collapse where welfare and wages are close to a subsistence minimum. This was likely the case in pre-industrial Europe, as for example, marriages responded negatively to increasing population size and decreasing welfare. Technological growth can be seen as an exogenous process which would gradually shift the line in figure 1 upwards, which would result in a period of population growth, followed by oscillation around a new higher equilibrium population size.

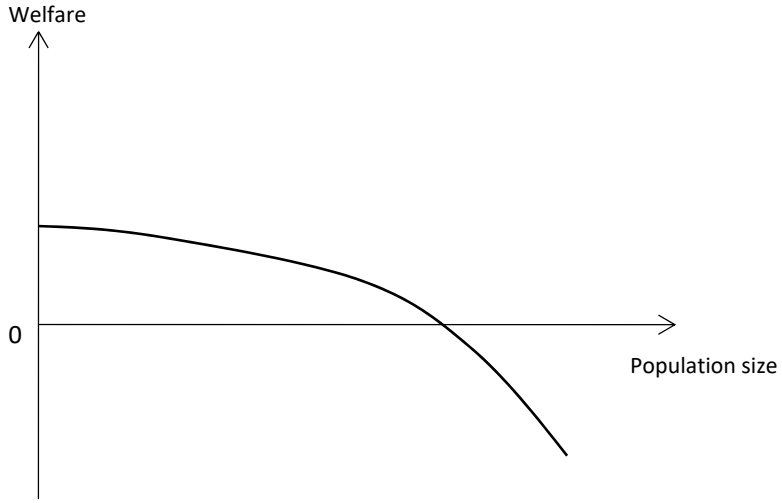


Figure 1: Malthusian relationship between population size and welfare

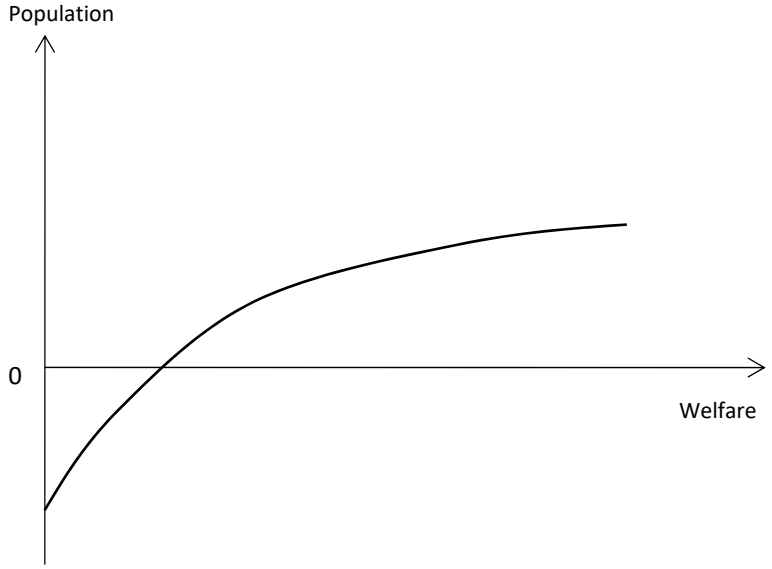


Figure 2: Malthusian relationship between welfare and population growth

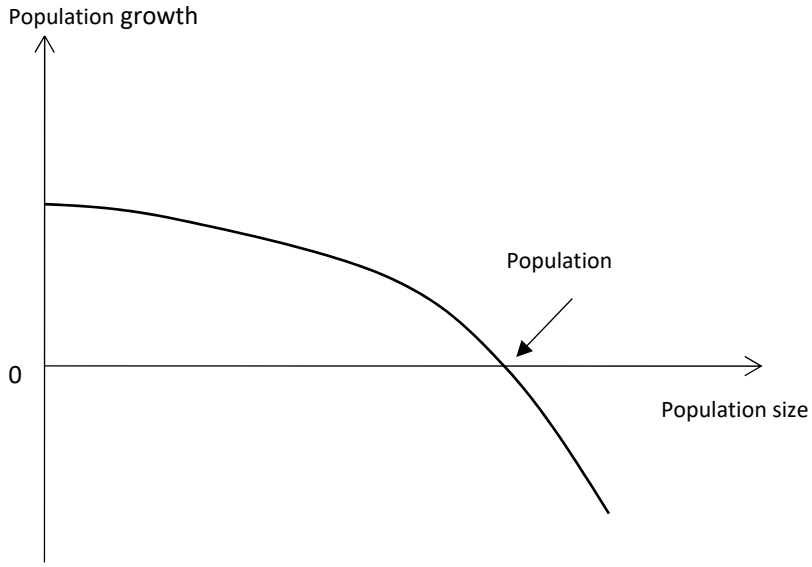


Figure 3: Relationship between population size and population growth in a Malthusian model

It is, as discussed early, likely that for some population levels an increasing population or higher population density could increase welfare. If such a relationship is true for any population size (as in figure 4) this would create a model with population size increasing towards infinity. If such a Boserupian relationship is dominant at lower population sizes (figure 4), but that ecological pressures would result in a negative relationship at higher population sizes (as in Figure 2), the relationship between population growth and population size would instead be roughly curvilinear as described by figure 5.[14] In figure 5 a combined Malthusian and Boserupian relationship is described at a given technological level, but it is possible that such processes would also increase the equilibrium population size over time. Boserupian processes such as increased intensification of land or adaptation spurred by increasing population density may stimulate technological growth, and allow higher and higher equilibrium population sizes. It is easy to construct such a model in which population size will grow forever. However, given the nature of

[14] It is not necessary to make any assumptions on the functional form of the relationships in any of figures 1-5 except that below some population level, an additional person increases welfare monotonically, and at above that population level there is instead a monotonic decline for each additional person. It also seems reasonable that population growth is above 0 for a very tiny population (though this is not necessary, there are examples on animal species that face extinction if population size fall below a small population size, a famous historical case is the North American passenger pigeon). There must also be some population level for which population growth turn negative, as the model would otherwise predict a population size that would move towards infinity.

exponential population growth, it is clear that a population cannot increase infinitely, and at some point each new individual must have a negative effect on the average welfare in a population. This implies that at some population size the relationship between population and welfare must turn negative also in such a model, and a relationship roughly like what is described in figure 5 should be accurate. As in a Malthusian model such a combined Boserupian-Malthusian model would have an equilibrium population size, at which the population would eventually oscillate around. However, unlike a Malthusian model, increasing population size would at lower population sizes increase population growth.

Demographers have empirically estimated how welfare is affected by population size, and how demographic rates, and consequently population growth, responds to changes in welfare. Economic demographers and historians examining European historical population have used wages in agriculture as a measure of welfare and resource availability, and found that when population increases this is associated with lower wages (as in figure 1). The elasticity of this effect seems to be about -1 (a 1% increase in population size, decreases wages by 1%), which is very substantial (Lee & Anderson, 2002). Social scientists have also examined how vital rates (i.e. population growth) have responded to changes in wages, primarily for fertility but also mortality. Researchers have consistently found that increasing prices and lower real wages reduce population growth (as in figure 2, e.g. Bengtsson, Campbell, & Lee, 2003; Lee, 1987; Tsuya, Feng, Alter, & Lee, 2010), though the magnitude of this relationship is weaker than for the effect of population size on wages. Similar relationships underpin ecological models on animals where similar strong effects between resources and population growth are observed (Lee, 1987; Sibly & Hone, 2002).

In preindustrial populations, it is clear that the relationship between population and welfare was Malthusian, and for most of human history it seems clear the human population size was regulated by welfare dependent homeostatic mechanisms (Lee, 1987). However, equally striking is the great, though gradual, increase in global population size over time. The rapid increase in population size during and after the industrial revolution is particularly striking. Axe planation for this observation is that technological growth is exogenous and that human innovation increased largely independent of population size or density. This increase in technology resulted in increasingly high productivity allowing higher population size. Such a relationship could be seen as exogenous shifts to the right in the relationship shown in figure 3, allowing higher and higher equilibrium population sizes. However, both theoretically and from empirical evidence (e.g. Boserup, 1965) it seems plausible that a growing population size itself could stimulate adoption of innovative behaviours increasing welfare.

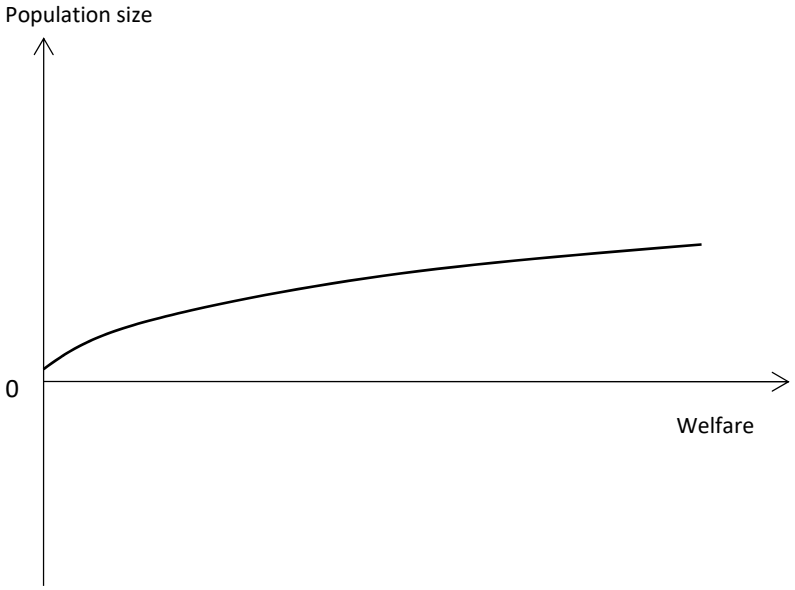


Figure 4: Boserupian relationship between population size and welfare

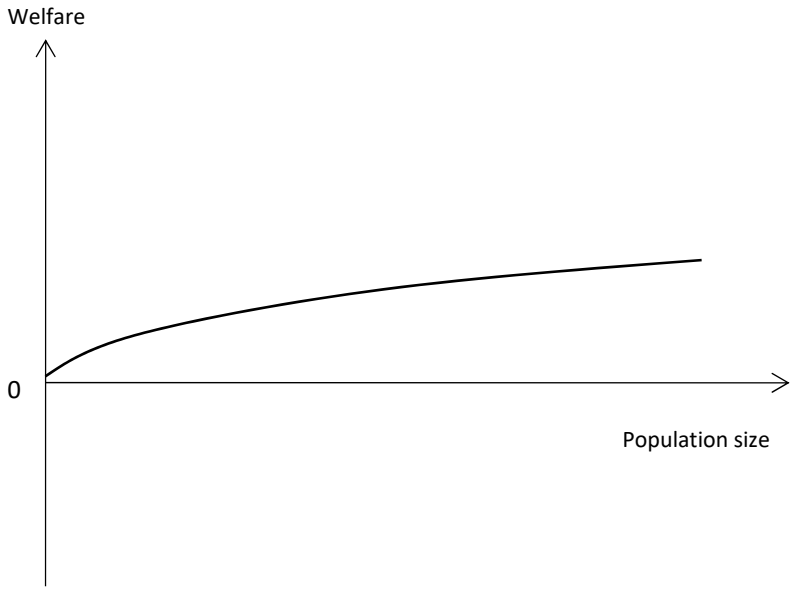


Figure 5: Combined Boserupian and Malthusian relationship between population growth and population size

If this is the case, technology, and in extension population growth, is under some circumstances positively associated with population size. The broad characteristics of such a relationship combined with a Malthusian relationship at higher population sizes were described in figure 5, but a number of researchers have created more sophisticated models examining such relationships.

Demographers have created analytical models, incorporating dynamic processes in which increasing population size increases welfare. Simple Boserupian models will result in both technology and population size both growing at increasingly rapid pace forever (e.g. the combination of figure 2 and 4) if there is no Malthusian mechanism. However as discussed earlier, population growth has to be bound by some factor eventually. More sophisticated analytical models combine Boserupian and Malthusian processes, and give further insights into how population size and growth interact endogenously and requires formal analytical notation or simulations (Chu & Tsai, 1998; Stefano Ghirlanda & Enquist, 2007; S. Ghirlanda et al., 2010; Lee, 1986; Pryor & Maurer, 1982; Simon, 1977). Such models include further assumptions than the very general relationship described in figure 5.[15] The relationship described in figure 5 has only two necessary assumptions; that for some population size (given abundant resources) there is a positive effect of a larger population to welfare, and that at some much higher population size this relationship must reverse.

Implications of Malthusian and Boserupian population models

Demographic theory on the relationship between population growth and population size has implications for how population ethicists think about the relationship between population size and welfare. If broad implicit assumptions such as a Malthusian relationship between increasing population size and welfare are used to formulate a scenario, it is important to understand why one makes such assumptions. Population ethicists commonly make thought experiments comparing populations of different sizes, with different levels of well-being. For such assumptions, on shape of the function between population size and wellbeing are relevant, which differently predict the effect of one additional person in a population on wellbeing. A common assumption in thought experiments is a Malthusian relationship between high population size and lower well-being. In many

[15] Examples of such additional assumptions in the models are: population sizes necessary to maintain technology, technological knowledge gradually lost over time, and technological growth is made possible from any surplus of resources above what is necessary for pure subsistence.

cases researchers also present a monotonic Malthusian relationship for any population size, and present examples with the highest welfare in populations with only a single individual.

A central conclusion from demographic theory, is that for very large population sizes an additional person will often decrease welfare. While this is not by necessity true at lower population sizes, it will probably always occur at some population size. The very broad relationships sketched in figures 3 and 5 can (at least in the long run) be interpreted as an universal aspect and feature of all biological life, and therefore something that theories on human behaviour should relate to. To analyse contemporary demographic and ethical issues in the present, such theory might give few guidelines to if the population of Earth is too small or too large, as the fundamental relationships described for contemporary societies are contested and many interpretations are plausible (see the following section).

It is important to distinguish between the concepts of optimal population size in population ethics, and equilibrium population sizes as presented here. The equilibrium population sizes do not maximize either average or total (or any other definition of) welfare in the population, unlike the concept of optimal population size. The optimal population size in both a Malthusian and a combined Malthusian/Boserupian model will be smaller than the equilibrium population size (no matter if the optimal population size is based on average or aggregated welfare). A Malthusian population equilibrium where the population size is close subsistence levels, is probably what Parfit (1984) had in mind when he described a thought experiment of a very large population with very low welfare as a setting for "the repugnant conclusion". The reproduced illustration for the repugnant conclusion of populations of different size with varying welfare (Parfit, 1984, p. 388) is the same as in figure 1 above.

The philosophical concept of welfare or happiness is different from an economic resource perspective on welfare, in which welfare is closely linked to surplus resources. It is plausible to distinguish a concept of a "life worth living" as discussed in population ethics, as different (and in income terms plausibly higher) than an individual with a salary just above subsistence minimum. However, non-material and material treatments of welfare are most likely interrelated, and many researchers assume that they to some extent overlap. A resource-based concept of welfare based on demographic theory suggests that the concept of a "life not worth living" is an unstable and short-term (though possibly recurring) deviation from a population equilibrium. According to demographic theory, population size will revert to intermediate equilibrium size in cases with very high and very low welfare. Adding people with very low welfare, will if a Malthusian model is applied, lower

welfare for everyone and at some point the demographic response of an additional individual would be negative future population growth.

The concept of unlimited reproductive rights gives a high priority to the free choice over reproduction also when significant (Malthusian) externalities exist. Demographic theory suggests that a population with high fertility preferences will move towards a situation with high population size, and low welfare. The equilibrium population size that results is higher, and the (both average and total) welfare is lower, than optimal population size derived from either a total or average view of welfare. Given unlimited reproductive rights, resource availability will regulate eventual population size.

Another implication of resource-dependent homeostatic models is that a population collapse, according to basic demographic or ecological models, results in strong population growth following the initial reduction in population size. A temporary exogenous change in population size is therefore recoverable from and temporary. Effects of population reduction on welfare are therefore also short-term. We also know of the human past that absolute extinction[16] is rare, consistent with demographic theory. Resource-dependent homeostatic models are central in basic ecological models of other species (Sibly & Hone, 2002), though there are also ecological examples of inverse relationships for low population densities (Courchamp, Clutton-Brock, & Grenfell, 1999). Population ethics of future populations are interested in the risk of total population extinction (e.g. Broome, 2010). If the welfare of future populations is not (or only slightly) discounted extinction incurs massive (sometimes close to infinite) welfare losses. However, if homeostatic population models are correct, population extinctions are highly unlikely, at least from gradual processes (e.g. climate change or environmental degradation), if not for sudden discrete catastrophic impacts if they kill an entire population at once (e.g. a gigantic asteroid). Homeostatic population models also have implications for intergenerational justice in population ethics. If population levels and welfare are both likely to revert to equilibrium levels, this may reduce intergenerational inequality for many outcomes, compared.

The concept of aggregated or total welfare is uncommon in population research with both ecological and economic assumptions. In most social science research, a strict "average" view of welfare is used. One reason for such perspectives is that an

[16] With absolute extinction, I mean that the population in an area declines to 0, through vital rates alone, without outmigration, cultural assimilation, or replacement by other populations in the same area. While dramatic, seemingly non-recoverable, population reductions are common, complete extinction appears rare. Famous example such as the Norse population on Greenland where abandoned through migration, and only a few (very inhospitable) Polynesian islands appear to ever have been abandoned, or even more speculatively population died out. Examples of such islands are the Auckland, Kermadec, Norfolk, and Pitcairns islands which all at some point were depopulated, though this likely took place through migration.

individual's resources are predictive of individual behaviour, and as such has explanatory power for a social system (for example the theoretical relationships described in this manuscript). Total welfare on the other hand does not correspond to a meaningful dimension to understanding why individuals in a society act the way they do. It has ethical implications vis-à-vis the concept of optimal population size for welfare, but is not very helpful in understanding the dynamics of individual actors and human population systems.

Populations change through population growth, which only indirectly affects population size, and this lagged effect has many implications. Changing circumstances will typically change mortality and fertility and through that population growth. This means that it is often misleading to look at associations between population size and welfare. This is particularly the case in a cyclical Malthusian model where the correlation between population size and economic wellbeing, differs fundamentally based on the strength of the two different assumptions of a Malthusian model (Lee, 1985). Changes in welfare, will affect population size through population growth, and that population size will respond only slowly.

Populations change through births and deaths, but births add very young people, while people that die are often older (or very young). As such both changes to the birth rate and the death rate will change the age structure, the proportion of individuals in different ages. Similarly, only some people in a population give birth (individuals that are neither very young nor very old), and the proportion of people in childbearing years will determine how many children are born. A consequence of this is that the proportion of young and old people are unlikely to ever be stable, even if in the long run birth rates and death rates are stable. In other words, all changes in vital events will also change the proportion of young and old individuals. Because humans are productive in the middle of their life and young and old individuals require more care, societies redistribute a large amount of resources across the life course in both pre-industrial and contemporary societies (Gurven & Kaplan, 2006; Lee & Mason, 2011). Therefore, an additional birth or death will also inevitably cause a change in the age structure of the population. Intergenerational justice issues related to age structure, and life course flows of resources, are central both to discussions of sustainability of pension systems in contemporary societies, as well as "obligations" or the value of more children in low fertility societies (e.g. Cigno & Werding, 2007; Gál, Vanhuysse, & Vargha, 2018).

What regulates population size in historical, modern, and future societies?

In previous sections, I gave a brief overview of how population size might interact with population growth, and technology at a theoretical level. I have also discussed some implications of such models. Below I will put the previous discussion in relation to empirical evidence from historical demography and demographic theory on how these relationships have changed over time.

Theorizing on population developments on pre-historic human societies universally assume that resource availability regulated population size. The very low global population growth rates the last 50,000 years, with the exception of the last two centuries, is strong evidence for such assumptions. However, we have close to no empirical evidence on actual population developments in the distant past, and as such the historical agricultural societies and dynamics of hunter-gatherer populations in the recent past has instead served as examples on how such relationships might have looked. Complicating such inquiries is that even populations that by all accounts must have relatively stable population size in the past, have at the time of contact with Europeans experienced explosive population growth (e.g. the Polynesians on Tikopia (Firth, 1936, 1965), and the !Kung people (Howell, 1979)).

There is more evidence from historical agricultural societies, where there is strong evidence for Malthusian population dynamics. Both studies using quantitative data relating population growth over time to economic variables (Galloway, 1988; Lee, 1987), as well as in-depth studies of various societal institutions in pre-industrial populations (Drixler, 2013; Firth, 1965) find that populations respond to resources in a Malthusian fashion.

Studies from pre-industrial Europe find that such relationships were substantial, but that they weakened considerably and disappeared during and after the industrial revolution (or at least were dwarfed by other societal changes). In order for a population to respond in a Malthusian homeostatic way to changing resources, population size must affect welfare (more people means lower wages), and welfare must affect population growth (higher wages means higher fertility and survival) (Lee, 1985). In pre-industrial Europe, the former effect was much stronger than the latter, but combined they were important enough to determine long-run population dynamics. As explanatory factors for changes in population in the short term, such dynamics had weak explanatory power, but still a consistent pressure downwards on population growth when population size was above its long-term trend (and vice-versa) was an inescapable determinant for long-term population developments (Lee, 1985). The unimportance for homeostatic tendencies as a short-term determinant relative to its still important role for long-term developments is important

to keep in mind when analysing the relationship between welfare and population growth in contemporary societies. Gradual technological development can either be interpreted as exogenous increases of equilibrium level between population and population growth, or as an endogenous process in which larger population sizes contributed to population growth. The demographic transition in itself is not really inconsistent with homeostatic resource regulation, as it essentially is a period of both rapid income growth and population growth, though at a shorter time scale there seems to be a less clear relationship between short-term income variation and demographic rates (Coale & Watkins, 1986).

Malthus wrote about different "checks" on population. Preventive checks regulate fertility and entry to sexual unions (age at marriage), and positive checks are related to catastrophic mortality and disease related to shortages of resources. The (Northwest) European marriage pattern has been described as Malthusian in that marriage ages dropped and spinsterhood decreased when economic resources were plentiful. This has been described as the most important way in which demographic rates were related to resources in for example pre-industrial England (Wrigley & Schofield, 1981). Mortality rates also increased when resources decreased. However, perhaps the most important determinant of population movements were epidemic diseases largely exogenous to population size, of which the Black Death in 14th century Eurasia is the most well-known case. The role of sudden and unpredictable mortality, and fertility which responded positively to a reduction in population size (and negatively to increases in population size) suggests that fertility was regulating populations in more predictable ways than mortality (Livi-Bacci, 2007). This is in contrast to a contemporary population where mortality is more predictable across years, while fertility cause more variation in population growth and age structure.

After and during the industrial revolution this traditional relationship between economic growth and changed dramatically. This can be interpreted that as that the rate of technological growth, or increase in living standards, for the first time outpaced population growth (though, population growth was still higher than ever). As both living standards and population growth increased simultaneously, and at an increasing pace until the 1960s, a broad homeostatic interpretation of this period is not possible. This is one reason that other factors than macro-level trends in welfare have dominated explanations to understand trends and determinants of mortality, fertility, and marriage, in the social sciences for contemporary societies. It is not possible to understand the fertility transition, the great decline in fertility in the 19th and 20th centuries in Malthusian terms, as welfare increased greatly at the same time as fertility fell. However, the fertility transition can be interpreted in broader homeostatic terms as a way for a population to regain population balance (and as a response to very high population growth).

In contemporary societies, the majority of the population on Earth lives in societies where the vast majority of mortality takes place in post-reproductive ages (United Nations Population Division, 2017). In such a context, fertility is the dominant determinant of population growth. Mortality after age 45 is not at all related to the NRR, or generational replacement. When mortality during and before reproductive ages are low, the difference in population growth between the fertility levels which maintains a stable population (just above 2 children per women), and levels of fertility in high-fertility societies are very substantial. Furthermore, in many populations in the world today the fertility rates are lower than replacement level fertility, and roughly half of the world population between 2010 and 2015 lived in countries with a Total Fertility Rate (TFR) that implied long-term population decline (United Nations Population Division, 2015). As such, fertility dynamics are clearly the major determinant to understand population dynamics in contemporary and future societies. From the individual perspective, it seems clear that people could, if they wanted, have more surviving children than they choose to have. This seems to be the case in the great majority of populations in the world today. This suggests that individual agency and childbearing preferences are the primary determinants for future population size. Individuals on average seem to at some level coordinate their fertility decisions with expectations on other aspects of lives, and on average choose fewer number of children as they judge that the economic, personal or societal costs of higher childbearing are undesirable. If global fertility levels will converge towards level above or below replacement levels seems critical for future population dynamics and the future of human population systems.

Overall, the world is still characterized by great variance in fertility, both across and within countries, and there have been great changes over time. With respect to welfare and fertility, there are many patterns that coexist at the same time that are in some aspects contrary to each other. In most societies, it appears that with respect to economic change and business cycles fertility at the population level has a pro-cyclical relationship today, similar to in the past (Sobotka, Skirbekk, & Philipov, 2011). At the same time, there is a robust cross-sectional association between low fertility and high wealth when comparing different fertility levels in different countries, though this relationship may disappear at very high levels of national wealth (Myrskylä, Kohler, & Billari, 2009). Such cross-sectional inferences from countries at different levels of development can be referred to as reading history sideways (Thornton, 2005), though such a broad relationship is also found when examining long-term time series for countries over time. Within pre-industrial societies, wealthy individuals traditionally have had higher fertility than poorer individuals have. This is partly reversed during industrialization and for a while many societies have seen negative status-fertility differentials between

groups (Skirbekk, 2008), though there is some evidence that some contemporary societies once again might see a positive wealth-fertility association at the within-society level (e.g. Andersson, 2000; Jalovaara et al., 2017). Overall, both in contemporary societies, and throughout the 20th-century female income seems more negatively related to fertility than male income.

How fertility will respond to welfare is critical for understanding the future of human population systems. At large population sizes, it seems inevitable that a broadly Malthusian link between welfare and population growth must exist, that at some population size will reduce population growth and cap population size. From what we know about human population systems and estimates of the carrying capacity of Earth this will most likely be achieved without catastrophic drastic population change, but rather through that humans will choose to have gradually fewer children when they face resource constraints (cf. Cohen, 1996), or in other words resource dependent homeostatic regulation. While a large number children today might be affordable given how much richer we are today than in the past, they are still very costly as a share of income of parents. Economists have suggested that the relative opportunity costs of children have increased with increasing investments in education and that this can explain declines in fertility (Becker & Lewis, 1974). If people eventually globally really desire fewer than replacement level children at current, or future higher levels of development, then a shrinking and not expanding population will be a more important determinant. Under such circumstances, Boserupian relationships will be more relevant than Malthusian relationships for the relation between welfare and population size.

Another dynamic, that like Malthusian and Boserupian welfare dependencies may have a positive effect on future population growth is if subsections of a population with higher than average fertility will have children that have more children than average. Such a mechanism will increase fertility everything else equal (Kolk, Cownden, & Enquist, 2014). If a cultural trait[17] of a parent both increases the number of individuals in the next generation (through higher fertility), and is acquired by those children, this will mean that the high fertility trait will be more common in each subsequent generation. This is true both if this is seen as a phenomenon arising due to shared socialization across children and parents (e.g. great love of children), if it is related to membership of sub-population with which is stable across generations with higher fertility (e.g. an ethnic or religious group), or even as a dynamic of entire populations where membership of that population is also acquired across generations (e.g. two countries with different fertility). In all cases, higher fertility individuals will increase as a share of the

[17] Any cultural or sociodemographic characteristic of an individual associated with fertility.

population over time. A genetic trait is also transmitted across generations and has the same dynamics (Burger & DeLong, 2016), though the process will be much slower than for cultural information. Such inherited fertility dynamics, make very low fertility preferences unstable (unless truly universal). The dynamics will be stronger when fertility is very low and many individuals have no children (and population growth negative). In fact, high pre-industrial fertility can be interpreted as a product of such processes maintaining fertility at a high level, where fertility will always be high enough to move towards a population equilibrium. Whether fertility decline was a result of a one-time unique circumstance or represents transformation that will continue to reoccur in highly developed societies is an open question (Kolk et al., 2014). In a world of low fertility preferences, such a dynamic will serve as a homeostatic mechanism. Similar to resource-dependent homeostatic mechanisms, the effect of inherited fertility may in the short term explain quite little of variance in fertility, while still have an important effect on long-term dynamics.

To solve the puzzles and explain current fertility variation within and across populations remains a challenge to contemporary demographers, though from these empirical observations it is still possible to draw inferences for the future. The most straightforward conclusion is that if fertility remains above replacement rate, the human population will at some point encounter resource limits that will bound human populations. Population growth in all societies has often at spells of time had population growth rates of at least 1%, given harsher socioeconomic circumstances than contemporary individuals face. Such population growth rates are impossible in the medium to long range, given the nature of exponential growth. Given future above replacement fertility levels, some homeostatic mechanism is therefore a necessity. We can also be largely certain that any such mechanism will be related to fertility, as any increases in mortality to substantively affect population growth will not occur.[18] We are also unlikely to run out of resources making human life possible (Cohen, 1996), so such regulations are almost certain to be welfare related as in a Malthusian model. In other words, upper Malthusian bounds exist on human populations which creates an upper bound population growth.[19] If humans will desire fewer children than what is necessary to maintain population size, minimum population sizes required for maintaining current technological levels will become a more important determinant for any relationship between population size and population growth. If highly developed future societies will tend to be character-

[18] For example, the great famines of the 20th century such as in China in the early 1960s and in Bengal in the 1940s had relatively little effect on long and medium term population developments.
[19] Even making wild hypothetical thought experiments such as interplanar travel does not change such boundaries fundamentally (Hardin, 1959)

rized by below replacement level fertility preferences, the amount of government-induced transfers from non-parents towards parents might be one of the most important determinants of future population size and growth.

Conclusions

In this manuscript, I described how there is both theoretical as well as empirical evidence to assume that population size and population growth are endogenously related. This is not surprising as the alternative would be that population change is random and that long-term population developments would be driven by a random walk. The broad theories relating welfare with population change might have weak short-term predictive power but are almost certain to be dominant over longer time perspectives. As population ethics often takes a long-term perspective understanding and engaging with demographic theory may be very productive. Even very general models on how population change might respond to welfare, would may change analytical models in population ethics substantially. The broad Malthusian and Boserupian relationships described in this manuscript are a useful starting point to consider such relationships. To make predictions over the long term for human future populations demographic theory is useful. It gives an overall framework to think about how population change and population size is related. Most uncertainty for future population developments are related to the historically completely novel phenomena of below replacement fertility.

Population ethicists may find a broader engagement and understanding of demographic theory useful from a number of perspectives. A more nuanced discussion of possible positive (Boserupian) as well as negative (Malthusian) effects of additional people on welfare is largely absent from the field. Strong predications from demographic theory may also help population ethicists evaluate different scenarios based on if they are empirically and theoretically plausible or implausible for human populations. This would be particularly useful when population ethics is applied to contemporary policy concerns. From a different perspective, many insights from population ethics are largely absent in demographic discourses. This is particularly true for perspectives on future generations and intergenerational issues, many of which are central in economic demography. Deeper engagement with demographers and population ethicists on these issues may give many new insights.

References

Andersson, G. (2000). The impact of labour-force participation on childbearing behaviour: pro-cyclical fertility in Sweden during the 1980s and the 1990s. *European Journal of Population, 16*(4), 293–333.

Arrhenius, G. (2011). The Impossibility of a Satisfactory Population Ethics. In E. N. Dzhafarov & L. Perry (Eds.), *Descriptive and Normative Approaches to Human Behavior* (pp. 1–26). Singapore: World Scientific.

Arrhenius, G., & Campbell, T. (2017). The Problem of Optimal Population Size". In M. Fleurbay (Ed.), *International Panel on Social Progress 1st Annual Report*.

Becker, G., & Lewis, H. G. (1974). Interaction between quantity and quality of children. In T. W. Schultz (Ed.), *Economics of the family: Marriage, children, and human capital* (pp. 81–90). Cambridge, MA: National Bureau of Economic Research.

Bengtsson, T., Campbell, C., & Lee, J. Z. (2003). *Life under pressure: Mortality and living standards in Europe and Asia, 1700–1900*. Cambridge, MA: The MIT Press.

Boserup, E. (1965). *The conditions of agricultural growth: the economics of agrarian change under population pressure*. London: Allen & Unwin.

Boserup, E. (1981). *Population and technology*. Oxford: Blackwell.

Broome, J. (1996). The welfare economics of population. *Oxford Economic Papers, 48*(2), 177–193.

Broome, J. (2010). The most important thing about climate change. In J. Boston, A. Bradstock, & D. Eng (Eds.), *Public Policy: Why Ethics Matters* (pp. 101–116): ANU E Press.

Burger, O., & DeLong, J. P. (2016). What if fertility decline is not permanent? The need for an evolutionarily informed approach to understanding low fertility. *Phil. Trans. R. Soc. B, 371*(1692), 20150157.

Charbonneau, H., Desjardins, B., Légaré, J., & Denis, H. (2000). The population of the St-Lawrence Valley, 1608–1760. In M. R. Haines & R. H. Steckel (Eds.), *A population history of North America* (pp. 99–142). Cambridge: Cambridge University Press.

Chu, C. C., & Tsai, Y.-C. (1998). Productivity, investment in infrastructure and population size: formalizing the theory of ester boserup *Increasing returns and economic analysis* (pp. 90–107): Springer.

Cigno, A., & Werding, M. (2007). *Children and pensions*. Cambridge, MA: MIT Press.

Coale, A. J., & Watkins, S. C. (1986). *The decline of fertility in Europe: the revised proceedings of a Conference on the Princeton European Fertility Project*. Princeton, N.J.: Princeton University Press.

Cohen, J. E. (1996). *How many people can the earth support?* New York: Norton.

Courchamp, F., Clutton-Brock, T., & Grenfell, B. (1999). Inverse density dependence and the Allee effect. *Trends in Ecology & Evolution, 14*(10), 405–410.

Dasgupta, P. (1998). Population, consumption and resources: ethical issues. *Ecological Economics, 24*(2), 139–152.

Drixler, F. F. (2013). *Mabiki: Infanticide and Population Growth in Eastern Japan, 1660–1950* Berkeley: University of California Press.

Firth, R. (1936). *We, the Tikopia*. London: George Allen & Unwin.

Firth, R. (1965). *Primitive Polynesian economy*. London.

Gál, R. I., Vanhuysse, P., & Vargha, L. (2018). Pro-elderly welfare states within child-oriented societies. *Journal of European Public Policy, 25*(6), 944–958.

Galloway, P. R. (1988). Basic patterns in annual variations in fertility, nuptiality, mortality, and prices in pre-industrial Europe. *Population Studies, 42*(2), 275–303.

Ghirlanda, S., & Enquist, M. (2007). Cumulative culture and explosive demographic transitions. *Quality & quantity, 41*(4), 591–600.

Ghirlanda, S., Enquist, M., & Perc, M. (2010). Sustainability of culture-driven population dynamics. *Theoretical population biology, 77*(3), 181–188. doi:DOI 10.1016/j.tpb.2010.01.004

Gurven, M., & Kaplan, H. (2006). Determinants of time allocation across the lifespan. *Human nature: an interdisciplinary biosocial perspective, 17*(1), 1.

Hardin, G. (1959). Interstellar migration and the population problem. *Journal of Heredity, 50*(2), 68–70.

Howell, N. (1979). *Demography of the Dobe !Kung*. New York: Acad. P.

Hutchinson, E. P. (1967). *The population debate: The development of conflicting theories up to 1900*. Boston: Houghton Mifflin.

Jalovaara, M., Neyer, G., Andersson, G., Dahlberg, J., Dommermuth, L., Fallesen, P., & Lappegård, T. (2017). Education, gender, and cohort fertility in the Nordic countries. *Stockholm Research Reports in Demography, 2017:06*.

Kolk, M., Cownden, D., & Enquist, M. (2014). Correlations in fertility across generations: can low fertility persist? *Proceedings of the Royal Society B: Biological Sciences, 281*(1779), 20132561.

Lee, R. (1985). Population homeostasis and English demographic history. *The Journal of Interdisciplinary History, 15*(4), 635–660.

Lee, R. (1986). Malthus and Boserup: a dynamic synthesis. In D. A. Coleman & R. S. Schofield (Eds.), *The State of population theory: forward from Malthus*. Oxford: Blackwell.

Lee, R. (1987). Population dynamics of humans and other animals. *Demography, 24*(4), 443–465.

Lee, R., & Anderson, M. (2002). Malthus in state space: Macro economic-demographic relations in English history, 1540 to 1870. *Journal of Population Economics, 15*(2), 195–220.

Lee, R., & Mason, A. (2011). *Population aging and the generational economy: a global perspective*. Cheltenham: Edward Elgar.

Lehmann, L., Aoki, K., & Feldman, M. W. (2011). On the number of independent cultural traits carried by individuals and populations. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*(1563), 424–435.

Livi-Bacci, M. (1999). *The population of Europe: a history*. Malden, Mass.: Blackwell.

Livi-Bacci, M. (2007). *A concise history of world population*. Oxford: Blackwell.

Myrskylä, M., Kohler, H.-P., & Billari, F. C. (2009). Advances in development reverse fertility declines. *Nature, 460*(7256), 741–743.

Parfit, D. (1984). *Reasons and persons*. Oxford: Clarendon Press.

Pryor, F. L., & Maurer, S. B. (1982). On induced economic change in precapitalist societies. *Journal of Development Economics, 10*(3), 325–353.

Sibly, R. M., & Hone, J. (2002). Population growth rate and its determinants: an overview. *Philosophical Transactions of the Royal Society B: Biological Sciences, 357*(1425), 1153–1170.

Simon, J. L. (1977). *The economics of population growth*. Princeton (N.J.). Princeton University Press.

Singer, P. (1976). A utilitarian population principle. In M. Bayles (Ed.), *Ethics and population* (pp. 81–99). Cambridge, Massachusetts: Schenkman.

Skirbekk, V. (2008). Fertility trends by social status. *Demographic Research, 18*(5), 145–180.

Sobotka, T., Skirbekk, V., & Philipov, D. (2011). Economic recession and fertility in the developed world. *Population and Development Review, 37*(2), 267–306.

Thornton, A. (2005). *Reading history sideways: the fallacy and enduring impact of the developmental paradigm on family life*. Chicago, Ill.: University of Chicago Press.

Tsuya, N. O., Feng, W., Alter, G., & Lee, J. Z. (Eds.). (2010). *Prudence and pressure*. Cambridge: MIT Press.

United Nations Population Division. (2015). *World Fertility Repord 2015 - data booklet*. New York: (United Nations publication).

United Nations Population Division. (2017). *World Mortality Report 2015*. New York: (United Nations publication).

Wilson, C., & Airey, P. (1999). How can a homeostatic perspective enhance demographic transition theory? *Population Studies, 53*(2), 117–128.

Wrigley, E. A., & Schofield, R. S. (1981). *The population history of England 1541–1871: a reconstruction*. London: Arnold.

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 519, SE-101 31
Stockholm, Sweden

Phone:
+46 8 402 12 00

E-mail:
info @iffs.se

Institute for FUTURES STUDIES