WORKING PAPER SERIES · Vol. 5 · 2023:1-11

STUDIES ON

CLIMATE ETHICS

AND FUTURE GENERATIONS

Working paper series 2023:1–11

Editors: Tim Campbell & Olle Torpman

Studies on Climate Ethics
and Future Generations
Vol. 5

Studies on Climate Ethics and Future Generations Vol. 5

Editors: Tim Campbell & Olle Torpman

Institute for Futures Studies
Working Papers 2023:1-11
Stockholm 2023

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/Tove Salomonsson
Cover image: Eirk van Hannen/Getty Images
Print: Elanders, 2023
Distribution: The Institute for Futures Studies, 2023

Contents

Preface

The Climate Ethics and Future Generations project has now completed its fifth year. It is hosted by the Institute for Futures Studies in Stockholm, and is generously financed by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences). The program is led by PI Gustaf Arrhenius, and co-PIs Krister Bykvist and Göran Duus-Otterström. It is interdisciplinary, involving philosophers, sociologists, economists and political scientists. The program aims to provide comprehensive and cutting-edge research on ethical questions concerning future generations in the context of climate change policy. It runs 2018–2023, and has now come to its final.

The project has three broad themes: *Foundational questions in population ethics*, which concerns how we should evaluate future scenarios in which the number of people, their welfare, and their identities may vary; *Climate justice*, which concerns the just distribution of the burdens and benefits of climate change and climate policy, both intra- and intergenerationally; and *From theory to practice*, which concerns how to apply normative theories to the circumstances of climate change, in light of both normative uncertainty and practical constraints. For more information about the program, visit climateethics.se.

The three themes are duly represented in this fifth volume of the program's preprint series, consisting of eleven papers in total. The first five represent the third theme: *From theory to practice*.

The volume's opening paper is the prize winning "How to feel about climate change?", co-authored by Julia Mosquera and Kirsti Jylhä. It examines the appropriateness of our different emotions in the face of climate change. The authors examine a number of normative criteria, and argue that what is appropriate to feel depends in part on the specific object of the feeling. It may be appropriate to feel anger about the emissions that cause climate change, while appropriate to feel appreciation of the warmer summers it causes. They also discuss instrumental reasons, e.g., to feel hope despite the unlikelihood of preventing climate change entirely. One key point of the paper is that a better understanding of the complex nature of climate emotions can help us handle the obstacles that our emotional disagreements pose for the cooperation that climate action requires.

The second paper, "How to value a person's life" by John Broome, compares two different strands in which economists typically value human life. On the first, which aims at being suitable for use in cost-benefit analysis, human life is valued on the basis of people's willingness to pay to reduce risks to their lives. On the second, which aims to be used for cost-effectiveness analysis in health care, human life is

valued on the basis of the length of the life and a measure of its quality. Broome argues that both strands are problematic, for either theoretical or practical reasons, but that they can nevertheless be reconciled into a method that has a defensible theoretical foundation and is of high practical relevance.

In the volume's third contribution, "DALYs and the Minimally Good Life", Tim Campbell discusses the related question of how to measure health impacts. The standard practice is to use Disability-Adjusted Life-Years (DALYs) as a proxy for health losses due to all disease causes, where one DALY represents the loss of the equivalent of one life year at full health. A natural thought is that governments should use their resources in such a way as to maximize expected DALYs averted per dollar spent. But this may conflict with a sufficientarian view that our priority should be to come as close as possible to a state in which every person can lead a life that is sufficiently good. Campbell's contribution draws attention to this conflict, and gives some reasons for being skeptical about sufficientarianism.

The volume's fourth paper, "Uncertainty Attitudes as Values in Science" by Joe Roussos, discusses whether and how science can be objective while realistically acknowledging and managing the impact of values in the production of scientific information. Although previous discussions have identified a great many locations where value judgements occur, they have focused on a particular kind of evaluation. The paper argues that philosophers interested in values and science ought to consider scientists' attitudes to uncertainty, which are evaluations of decision situations. Roussos' main claim is that if you are concerned about inductive risk in a particular part of science, then that concern should include uncertainty attitudes alongside the more commonly considered moral, social, or political values.

In "Sex Selection for Daughters", Karim Jebari and Martin Kolk discuss an issue related to population size: sex selection. Sex selection reflecting a preference for male children is a well-known ethical issue, but Jebari and Kolk focus instead on the newly emerging issue of sex selection reflecting a preference for female children. Current technologies make it possible for parents to choose the sex of their offspring. Jebari and Kolk consider the possible demographic consequences of the adoption of these technologies, in light of an emerging trend of parents preferring female children. They argue that female-biased populations are likely to grow faster than male-biased populations and populations with an equal sex ratio; and they predict that cultural norms promoting female-biased populations will be self-reinforcing.

The next two papers represent the second theme of the research program: *Climate Justice*. In the first of these papers, "Inducement-Based Emissions Accounting", Olle Torpman argues that a just distribution of climate burdens must be based at least in part on considerations of who has emitted how much. In order to

settle that, he argues that a reliable emissions-accounting method is needed. Torpman moreover argues that existing emissions-accounting methods fail to identify the appropriate responsible-making feature of agents, and that they are therefore implausible. He proposes a new emissions accounting method – *inducement-based emissions accounting* – aimed at avoiding the problems faced by the current methods.

In the second paper on climate justice, "The Ethical Risks of an Intergenerational World Climate Bank", Stephen Gardiner delivers some criticisms against the idea of a World Climate Bank, that has been proposed by John Broome and Duncan Foley, among others. Gardiner argues not only that there are certain injustices built into such an institution that must be addressed, but also that there are a variety of problems with the assumptions that underlie the argument in favor of such a solution to climate change. This, Gardiner argues, also affects the pragmatic relevance of the idea. Instead of a world climate bank that shifts all the burdens of climate action to the future, Gardiner proposes a Climate Justice World Bank that respects principles of global and intergenerational ethics.

The volume's final set of papers are the most theoretical, all representing the first theme of the program: *Foundational issues*. Katie Steele's "Longtermism and Neutrality about More Lives" is an important contribution to a current debate about whether we have moral reason to reduce the risk of futuristic threats to humanity's survival, even if doing so would involve considerable opportunity costs for present people. Some argue that the claim that we have such moral reason depends on a kind of totalist utilitarianism; and they argue that we should reject totalism in favor of the view that the addition of worthwhile lives to the world does not make the world better or worse, but is inherently *neutral*. Steele shows that these philosophers are mistaken, since conclusions about futuristic threats of any kind are not greatly dependent on totalist utilitarianism, or on what she calls "totalist moral mathematics". According to Steele, our predictions about those who may or may not exist later inevitably matter for the purpose of determining the strength of our moral reasons to act in ways that will affect the present generation.

Next, in "Population, Existence, and Incommensurability", Melinda A. Roberts defends an approach to population ethics that captures the intuition that leaving a person out of existence altogether doesn't make things morally worse. Her guiding idea is *person-affecting*: you can make things morally worse only if you make things worse for a person who does or will exist. She gives a precise statement of the principle she thinks captures this central intuition, and asks whether there is a tension between this principle and certain theoretical assumptions, such as Transitivity–that the relation 'better than' is transitive–and what she calls Trichotomy–that the relations 'better than', 'worse than' and 'as good as' exhaust the possible value relations. Roberts argues that her approach is compatible with both assump-

tions, and that therefore the person-affecting idea requires neither value incommensurability nor betterness cycles.

In the volume's tenth paper, "Do We Owe the Past a Future?", SJ Beard and Patrick Kaczmarek defend a view they presented in a previous paper. This view consists primarily of two claims: (i) that by preventing human extinction, we can render the sacrifices that past people made to benefit those who would come after them more worthwhile, and (ii) that by squandering their sacrifices, we wrong these past people. In particular, the paper responds to some criticisms that Elizabeth Finneron-Burns has raised to these claims.

In the volume's eleventh and final paper, "Scanlonian Contractualism and Future Generations", Emil Andersson, Gustaf Arrhenius, and Tim Campbell consider different problems in population ethics from the point of view of Scanlonian Contractualism. There are features of this view that make it difficult for its proponents to reach seemingly obvious conclusions in cases where we are to choose which population to bring about. The authors discuss different interpretations of Rahul Kumar's idea of "standpoints", but argue that it runs into trouble with regard to the individuation of standpoints. Scanlonian Contractualism cannot, it seems, avoid the aggregation problems that standard "impersonal" theories face without running into other problems that are even worse.

We are pleased to be able to share this new work from the Climate Ethics and Future Generations project. As with previous volumes, the authors of these papers would greatly appreciate any comments, questions, and objections that you wish to share with them. Contact information is found at the front page of each paper. We would also like to thank Gustav Hedlund for helping out with formatting most of the papers in this volume.

Tim Campbell & Olle Torpman
Editors

Julia Mosquera[1] & Kirsti Jylhä[2]

How to Feel About Climate Change? An Analysis of the Normativity of Climate Emotions[3]

Climate change evokes different emotions in people. Recently, climate emotions have become a matter of normative scrutiny in the public debate. This phenomenon, which we refer to as the normativization of climate emotions, manifests at two levels. At the individual level, people are faced with affective dilemmas, situations where they are genuinely uncertain about what is the right way to feel in the face of climate change. At the collective level, the public debate reflects disagreement about which climate emotions are appropriate to feel. The aim of this paper is to examine the normative reasons in favour of different climate emotions by combining normative criteria from philosophy and psychology. We conclude that these criteria provide partial reasons for or against different climate. Emotional disagreement in climate contexts may generate

[1] IFFS, julia.mosquera@iffs.se

[2] IFFS, kirsti.jylha@iffs.se

[3] A version of this paper is published in the *International Journal of Philosophical Studies*: Julia Mosquera & Kirsti M. Jylhä (2022) How to Feel About Climate Change? An Analysis of the Normativity of Climate Emotions, 30:3, 357–380, DOI: 10.1080/09672559.2022.2125150. This paper is the winner of the 2021 PERITIA Prize on the social and political significance of emotional attitudes and responses, part of the 2021 IJPS Robert Papazian essay competition on Ethics and Emotions and funded by the UCD Centre for Ethics in Public Life. The winning essays reflect some of the main themes and interests of the project Policy, Expertise and Trust in Action (PERITIA), funded by the European Union's Horizon 2020 research and innovation programme (grant No 870883). We are grateful to Krister Bykvist, Maria Ojala, and Tim Campbell for providing valuable comments to an earlier version of this paper.

distrust, potentially hindering cooperation for climate action. We propose that we can ease challenges like this if we come to terms with the complex nature of climate emotions and their normative justification.

1. Introduction

Climate change has become one of the most emotionally loaded issues of all times. Given the urgent nature of the climate crisis, worry and concern over the issue has become evident. The latest UNDP report warns that climate change is a contributing factor to the current state of "anguish" in which humanity finds itself, where 6 in 7 people worldwide are plagued by feelings of insecurity (UNDP, 2022). Alarmed feelings are widespread in response to the climate crisis, ranging from milder and adaptive forms of emotions to more pathological states that can include serious mental health problems and functional impairment. Yet, climate change does not evoke concern in everyone. Some feel indifferent or even doubt the existence of human-induced climate change and may even feel irritated over the frequent debates and emotional displays of others.

Interestingly, climate change emotions have recently become a matter of normative scrutiny. Questions about how oneself or others *ought* to feel in the face of climate change (as opposed to how we and others *actually* feel) have recently gained attention. We refer to this phenomenon as the *normativization of climate emotions*. The normative practice of the appropriateness of climate emotions manifests at various levels. At the individual level, people find themselves confronted with genuine *climate affective dilemmas*. Can I feel happy and enjoy increasingly hot summer days, or should I instead feel and display worry, given the source of this warming? Is it legitimate to feel hope and optimism in the face of catastrophic scenarios?

At the interpersonal level, the normativization has become particularly visible in the public domain where there is evidence of an ongoing process of negotiation of the normative status of climate emotions. This negotiation manifests in the criticism or blame towards others' emotions for failing to conform to certain standards of appropriateness (e.g., feeling 'too scared' or feeling 'too relaxed' about the threat of climate change), as well as attempts to elicit, provoke, or induce specific climate related emotional responses in others through harsh public messages of indignation. We refer to these processes as *climate affective disagreement*. Additional questions regarding the climate attitudes of others arise here. Should climate anxiety be considered an overreaction and a sign of mental problems and alarmism, or a desirable and/or rational response to a crisis? Are there reasons that suggest that some emotions are wrong, inappropriate, or counterproductive? And if some emotions are more appropriate than others in response to climate change, are we justified in correcting

and blaming those who don't display them? As the climate crisis continues, new emotional reactions, emotional disagreements and affective dilemmas may emerge.

The aim of this paper is to understand the normative considerations that justify the role of different emotional responses evoked by climate change, and to evaluate the appropriateness and usefulness of these emotions. While some of them might a priori seem inappropriate (e.g., the enjoyment of warmer summer days), they might still be appropriate from the point of view of rationality. The opposite might be true of other emotions, which might not be particularly fitting in terms of rationality, but appropriate from the point of psychological human predisposition and wellbeing, and useful in achieving climate action (e.g., hope in the face of climate threat). For this purpose, we combine normative approaches from philosophy and psychology. While philosophy possesses long-standing, sophisticated accounts to analyse the normativity of emotions, these are rarely applied to specific emotional settings, including climate emotions. And while the existing psychological research on climate emotions provides developed taxonomies and measures of climate emotions, it usually ignores the normative aspects of those emotions, or takes certain normative assumptions for granted. Given the fundamental role of emotions in our everyday social practices and public life, insights regarding the normativity of climate emotions are crucial. We argue that this can improve our understanding of the complexities behind discourses on climate emotion, and promote social trust, collective action and civic and democratic practices aimed at addressing climate change, and a responsible approach to evoking climate emotions in in media and society.

We approach climate change as an umbrella term that includes several different types of phenomena. It includes physical events (e.g., the rising of the average temperature on Earth, the melting of the glaciers) that do not entail evaluative descriptions. Other elements incorporate evaluations on what climate change directly implies to life on earth (e.g., climate change as a threat to human life); on what is being done to address climate change at system-level (e.g., political decision-making) and individual level (e.g., stop flying). And others are mostly of an evaluative nature regarding practices and behaviours related to climate change (e.g., the unjust distribution of burdens and benefits of the effects of climate change, free-riding practices, political inaction). The different nature of these objects calls for different types of affective responses (e.g., fear, anxiety, anger) which will be measured against different standards of appropriateness.

The paper proceeds as follows. In section 2 we present and analyse some paradigmatic examples of the negotiation of climate emotions present in the public domain. We show how the public engages in the normativization of climate emotions, identify some of the different standards they rely on when formulating their criticisms and suggest different reasons for why they may rely on those standards. In section

3, we present some criteria of evaluation well-known in the philosophical literature on emotions, namely fittingness, warrant, and prudential considerations. We apply these criteria to some paradigmatic climate emotions and emotional dilemmas individuals increasingly face in the current climate scenario with the aim of understanding the capacity of the different emotions to represent the objects they are directed to. In section 4, we present an analysis of the psychology of climate emotions and discuss the implicit and explicit criteria for evaluating appropriateness of emotions and the psychological rationale for their use. Section 5 is the conclusion, where we provide insights and recommendations for further research.

2. Affective normativity in the climate change public debate

The normative practice of emotional appropriateness is present in different spheres of our everyday life. We engage in this practice when we give and ask for reasons to respond emotionally to properties and objects in an appropriate manner. For this purpose, we create and participate in various sub-practices that facilitate the mutual correction of people' emotional responses and the collaborative discovery of evaluative properties (Gallegos, 2021).

This practice has become particularly salient in the context of climate emotions. This is exemplified by the apparent disagreement existing in the public domain regarding which emotions ought to be endorsed in the face of climate change and the related emotional attitudes displayed as part of this disagreement, including criticism, blame, or attempts to elicit certain emotions. These types of criticism don't usually invoke explicit appeal to specific normative standards from the philosophy or psychology of emotions. However, the negotiation seen in the public domain does rely on some sort of basic norms or standards against which people's climate emotions are compared.

The public exchanges between the Swedish climate activist Greta Thunberg, the former president of the USA, Donald J. Trump, and the president of Russia, Vladimir Putin, are paradigmatic examples of the normative practice taking place in the public debate. The most well-known public speeches of Thunberg convey her frustration, anger, fear, and the rejection of hope. For example, during her participation at the World Economic Forum in Davos in 2019, Thunberg said (World Economic Forum, January 25, 2019):

> I don't want your hope. I don't want you to be hopeful. I want you to panic. I want you to feel the fear I feel every day, and then I want you to act. I want you to act as you would in a crisis. I want you to act as if our house is on fire. Because it is.

Thunberg has been addressed by both Trump and Putin. In response to her emotional speeches, Trump wrote the following ironic tweet: *'She seems like a very happy young girl looking forward to a bright and wonderful future. So nice to see!'* (Trump, 2019). And after Thunberg was named Time's Person of the Year in 2019, Trump wrote another tweet, referring to her pleas for governments to stop global warming: *'So ridiculous. Greta must work on her Anger Management problem, then go to a good old-fashioned movie with a friend! Chill Greta, Chill!'* (Trump, 2019).

Putin made a somewhat similar critical assessment about Thunberg's campaign in 2019 at an energy forum in Moscow, by first voicing seemingly benevolent intentions (C.f., BBC News, October 3, 2019 and Bloomberg Quicktake, Twitter Post, Oct 3 2019, 12.05 AM):

> You know, young people, teenagers, draw attention to today's acute problems, including environmental problems, and it is right, it is very good. They definitely must be supported. But when someone uses children and teenagers for someone's benefit, it is only reprehensible ... I am sure that Greta [Thunberg] is a kind girl and very sincere. But adults must do everything to not put teenagers and children in extreme situations, they must shield them from extreme emotions that could destroy a personality... Nobody explained to Greta that the modern world is complicated and complex, it changes fast. People in Africa and in many Asian countries want to be as wealthy as people in Sweden. How can it be done? By making them use solar energy, which is plentiful in Africa? Has anyone explained the cost of it?

Exchanges like this are particularly relevant here given the different but eminent role that each of these three actors has in the climate change context. Thunberg is a highly influential climate activist with a proven capacity to mobilize a great part of the young population. Trump and Putin are political leaders of two nations with high responsibility for past emissions and whose political involvement is crucial in climate mitigation policy. Furthermore, important fractions of the population take these agents, their political institutions, and their decision-making as worthy of social and political trust. This means that public statements like these have the capacity to attract great attention from voters and supporters, who might take these statements as expressing truths about what are the corrective affective responses to climate change.

There are interesting similarities between the rhetorical elements displayed in Tump's and Putin's remarks. Both criticised Thunberg's words, dismissed her message, and implied that her fear was unwarranted. Putin disregarded Thunberg's concerns and dismantled her authority by claiming that she (and as an extension,

her conclusions and emotions) have been manipulated by other, more capable persons, a common line of criticism towards Thunberg among the public.

One interpretation is that these comments respond to an intentional mechanism to shift away the focus of the public discussion from the object of Thunberg's emotional attitude, namely the political inaction in the face of the potentially devastating effects of climate change, to the emotional component of her reaction. This move can be interpreted as an *appeal to emotion* fallacy, where the appeal to emotion is used to either convince or discredit a discoursal opponent, or a or *'tone policing'* fallacy, an *ad hominem* type of fallacy where the tone of the speech is used as a reason to discredit the validity of the content of a message and sometimes also of the interlocutor, like when a woman is told to "be less emotional" after expressing a concern that affects her emotional state, or a black activist is told to "calm down" when expressing discomfort about an injustice with an angry tone.[4]

Often, this shift away from the content of a message is combined with the consideration of emotions and their display in certain contexts as evidence for the mental state or mental health of the subject, often erroneously but intentionally. This is particularly salient when emotions are displayed by women. In our society rationality has traditionally been considered as superior (and stereotypically male) and emotions and the expression of emotions as inferior (and stereotypically female) (Damasio, 2005; Plant, et al., 2000), and this could influence the public perceptions on aptness. This could in part explain why it has been possible to depict climate anxiety as hysteria (sometimes directly in relation to female gender), and the non-emotional approach as a rational and objective response supported by healthy skepticism (Pettersson et al., 2022; Toivonen, 2022), despite the strong evidence showing the severity of the climate crisis.

In the case of both Putin's and Trump's comments, their rhetoric could also be analysed in the context of a power struggle and hierarchy. Particularly Putin's remarks echoed societal perceptions regarding minors and women as being vulnerable and innocent and, by extension, someone who should be protected and cherished by adults and men, respectively. Such views could be interpreted as manifestation of a seemingly benevolent form of sexism that allow positive or even admiring views on women while not guaranteeing them actual influence in society (Glick & Fiske, 2001). There are also paternalist attitudes, echoed in the tendency to ignore the worries and the demands of the young, while depicting them as ignorant children whose place is not in demonstration but in school to learn more (Bergmann & Osse-

[4] For an account of 'tone policing', see Ijeoma Oluo's *So you want to talk about race* (2018). Her reconstruction of the use of this fallacy is rooted in the observed relations of power, privilege and racism existing behind the silencing practices toward black rights activists. We expand the philosophical analysis on the dismissal of fitting climate anger in Section 3 below.

waarde, 2020). Accordingly, depicting Thunberg's reaction as misguided and emotional (and hence, a reaction that makes her an object of protection instead of an authority) could be a rhetorical move, but can also reflect how she is *actually* perceived by some merely due to her age and gender. Consequently, her message and emotions could be dismissed regardless of their content and character.

In sum, it has become evident that we are not indifferent towards our and other's emotional responses to climate change-related events, as well as that we rely on different types of evaluative criteria to judge the appropriateness of these emotions. For the best of our knowledge, the criteria employed to evaluate climate emotions have not been subject of systematic research in philosophy and psychology. Below, we provide an overview of these criteria with the aim of improving the understanding of the complex nature of climate emotions, their normative justification, and their role in trust and collective action, much needed to solve the climate crisis.

3. Philosophical normativity and climate emotions

Within philosophy, there is wide acceptance of the idea that emotions can be appropriate or inappropriate. Criteria of correctness vary depending on the view one holds about the nature and character of emotions. A paradigmatic view of the nature of emotions is that emotions are evaluative representations of formal objects of the world that contain value-laden features (D'Arms & Jacobson, 2000; Greenspan 1988; Roberts, 1988; Solomon, 1976). Under this view, emotions are somewhat analogous to beliefs and thus they can be assessed in terms of their cognitive rationality. This includes their fittingness or aptitude to represent the properties instantiated by the objects towards which emotions are directed, their warrant and coherence, or their capacity to relate to other evidence-sensitive evaluative processes. According to the fittingness criterion of emotions, an emotion is rational in terms of being "fitting", "correct", or "appropriate" if there is a representational match between the emotion and the object toward which the emotion is directed. For example, fear is fitting in those situations in which it is directed towards objects that are genuinely dangerous, since fear is a representation of danger.

An important caveat to this account is that the fittingness or appropriateness of an emotion is not meant to be a moral evaluation of it. Just as beliefs can be true or correct when their representational content matches the world and their objects, so can emotions be fitting or correct when they appropriately match the properties instantiated by their objects. Thus, the question of *correctness* is different from the question of whether a feeling or an emotion is morally permissible or is "what to feel" all things considered. Offensive, immoral art or jokes are usually provided as examples where it can be fitting to feel pleasure and amusement, respectively,

despite their immoral character (Jacobson, 1997). The wrongness or viciousness of a joke isn't itself a reason for a joke not to be amusing. Thinking otherwise commits us to the so-called "moralistic fallacy" that some have warned against (D'Arms & Jacobson, 2000). The idea is that there are different grounds to evaluate jokes, and although offensiveness might be a property of these jokes, it isn't itself a *relevant* reason for a joke failing in the domain of amusement; doing so would be appealing to the "wrong kinds of reasons" (McDowell, 1987). Reasons of the right kind for one to hold a certain attitude, the argument would go, are those that bear on certain properties of the object—for instance, reasons to be amused by a joke that bear on whether the joke is particularly amusing.[5]

Fittingness is behind some of our practices of criticism of emotions (e.g., "Don't be sad—it's not such a big deal") and it is often invoked when reflecting on how we should feel. Proponents of this account stress its usefulness over competing normative criteria: "(...) Prudential considerations, especially about fear or anxiety, are often counterproductive; and moral considerations can induce guilt without alleviating the offending emotion" (D'Arms & Jacobson, 2000: 73). Avoidance of guilt and contribution to productive action are features that make fittingness a particularly interesting normative tool for the analysis of climate emotions.

Sweden and several other parts of the world, including other European countries, Canada, and Australia, have experienced exceptionally warm summers since 2018 (Painter et al., 2021; Wilcke et al. 2020). This is an important climate change-related phenomenon. It is one of the closest and most tangible effects of climate change that those living in mild regions of the globe can relate to in their everyday experience.[6] During exceptionally warm days, people have started to feel genuinely torn between the enjoyment of sunny and warm temperatures and the feeling of guilt or uneasiness provoked by the realization of the causal connection between climate change and warmer days. This phenomenon has been echoed in national and international media, with headlines that include explicit reference to the state of cluelessness in which we find ourselves regarding the appropriate affective attitude to have toward increasingly warmer days.[7] This is a paradigmatic example of what we call *climate affective dilemmas*, the type of affective dilemmas that arise

[5] The challenge here is how to explain this evaluative relationship in a non-circular manner. For a thorough discussion of this challenge see Bykvist (2009).

[6] We focus on warmer summer days for simplicity, although something akin could be argued of warmer winter days.

[7] Some examples of the national and international coverage of this phenomenon include: "Can we enjoy, or should we be ashamed?" (*Göteborgs-Posten*, 2019), "How long can we call the heat 'nice weather'?" (*Svenska Dagbladet*, 2020)) and international ("Am I the only one who's terrified about the warm weather? *(The Guardian*, 2019), "Is it okay to enjoy the warm winters of climate change?" (*The Atlantic*, 2017), "You can care about climate change and still enjoy freakishly warm winter days" (*The Washington Post*, 2017).

as a conflict between different emotions elicited by climate change and other related phenomena.

By *affective dilemmas* we refer more generally to situations in which individuals face a conflict between two or more incompatible emotional responses to an object or phenomenon where there does not seem to be an obvious solution as to how one overall ought to feel in the face of it and where the emotional status quo is not an option. In affective dilemmas, the overall most fitting attitude can be one of ambivalence, which does not have a counterpart in the case of traditional act dilemmas.[8] Affective dilemmas are normative in the sense that they entail the existence and comparison of different types of normative reasons in favour of the different possible affective responses one is faced with.

In the face of increasingly warmer summer days, there are prima facie reasons to feel both anxiety and joy. On one hand, there are reasons to feel anxiety about warmer summers because these are goods, such as the enjoyment of the warm weather, that are morally tinted by their causal source, namely humanly caused climate change with potentially devastating consequences for life on the planet.[9] There are culture specific norms that regulate the display of emotions and in some social and cultural contexts it seems to have become an implicit social norm not to express joy regarding exceptionally warmer summers; and if joy is expressed, it is expected to be usually accompanied by a reference to how 'weird' or 'scary' this warmth feels. On the other hand, there are also reasons to feel joy during warmer summer days given that warmer weather is pleasurable—a preference shown to be extended among most people (Pew Research Center, 2009).[10]

It could be suggested that failing to feel anxious or fearful in the face of increasingly warmer summer days is somehow contrary to virtue. It could be that the reasons for feeling anxiety or fear in the face of increasingly warmer summer days are stronger or of a special nature, and thus in this scenario defeat the reasons to feel joy. Warmer weather is known to be caused by climate change. Failing to feel anxious or fear would seem to show some sort of objective irresponsiveness to the

[8] One may argue that affective dilemmas of the sort described here can be reduced to traditional moral act dilemmas. In this way, the conflict individuals are presented with is simply a conflict between whether to go out and enjoy the sun or whether to stay at home, and thus are a conflict between two actions both of which seem equally permissible, but where only one can be chosen (where the status quo is not an option). The type of dilemmas we present here differ from traditional moral dilemmas and thus cannot be subsumed to those. In affective dilemmas, what individuals are primarily uncertain about what is what is permissible or appropriate to feel in a given situation, and thus face a situation of what we could refer to as 'affective uncertainty'. The nature of this affective uncertainty is partly epistemic and partly moral.

[9] For a thorough discussion of this phenomenon, see Cullity's *Concern, Respect, and Cooperation* (Part II: 'Moral Derivations', pp. 67–172).

[10] It should be though mentioned that even in milder regions, temperatures currently reach levels that can cause physical health problems and strong discomfort among some parts of the population.

cause of warmer summers, climate change, and its dangerousness for life on this planet. So, if failing to feel anxious or fearful in the face of increasingly warmer summers is inappropriate because it is irrational (in terms of irresponsiveness to reasons), this would seem to solve the affective dilemma we presented above.

There is however a further question as to what is the object that our affective attitudes are directed towards is in this context. Emotions are sometimes understood as affective reactions directed towards 'simple' objects (e.g., warmer temperatures) or bundles of objects with different metaphysical and causal properties (e.g., climate change), which would include features like their causal properties. However, individuating the different objects of emotions is important, particularly in the case of climate change. The climate emotions literature rarely provides a systematic distinction between the different objects of those emotions, despite climate change being known to be a complex phenomenon composed by objects of different nature. This is an obstacle for understanding the role of climate emotions, for evaluating and for comparing them.

'Being caused by human-induced climate change' is a property of increasingly warmer summers and it is reasonable to expect that many of us believe this to be the case. Thus, fear can be an appropriate response to this feature of warmer summers. However, these grounds seem irrelevant for judging warmer summer days in the domain of pleasurableness and thus a relevant reason for summer days failing at being pleasurable. In this domain, the relevant object of emotion is the warmth or the experiential part of warmer summers, toward which it would just be fitting to feel pleasure or joy. Doing otherwise would be an instance of the 'wrong kind of reasons' explained above.[11]

There are other instances where clarifying the object of a climate emotion can help understanding the appropriateness of different emotional reactions. Climate hope is widely discussed among psychologists of climate emotions, and some have

[11] For a contrasting approach to the structure of reasons, see Dancy's *Moral Reasons* (1993). In his account, relevant reasons cannot be atomised in the way done above. The causal etiology of an event, i.e., the fact that e.g., it is achieved through immoral means, can undermine or 'disable' the event from being reason-given. Think of sadistic pleasure. The fact that a pleasurable activity is sadistic disables pleasure in this context from being reason-giving. So, there is no further reason against the promotion of pleasurable sadistic activities that 'outweighs' the reason based on the pleasure, because there is no pleasure-based reason in the first place. However, had the pleasurable activity been an innocent one, the fact that is pleasurable would have been reason-giving. The same kind of reasoning can be suggested to apply to warmer summers caused by climate change. The fact that warmer summers are caused by climate change disables, so the argument goes, the pleasurableness of the warmth from being reason-giving, so there is no further reason that outweighs the pleasure-based reason, since there is no pleasure-based reason in the first place. We thank an anonymous reviewer for raising this analogy. While we might agree that sadism, if intrinsically bad, can disable the pleasure from sadistic pleasurable activities from being reason-giving, it is less clear that climate change, both instrumentally bad and good, can disable warming temperatures resulting from climate change from being reason-giving in terms of pleasure.

recently emphasised the importance of this emotion for dealing with climate change (Geiger et al., 2021; Ojala, 2012b). In the case of climate hope, one may wonder about its appropriateness in terms of the fittingness criterion introduced above. According to this criterion, emotions can be correct or incorrect depending on how well they represent the properties of the object at which they are directed. Climate hope, where hope is understood in a 'narrow' sense, i.e., as a response directed to a mental state of the kind 'I have hope that we will fix climate change in x amount of time', could be deemed as unfitting under this account, given the low odds of success in fixing climate change in that given amount of time, thus not matching the current state of affairs with respect to climate change. However, hope becomes more fitting if it is understood as a general emotional state of the form 'I am hopeful that climate change will be fixed despite of the high degree of uncertainty that we will succeed'. A similar understanding of hope is behind recent proposals like "hopeful pessimism", an account of pessimism that rejects 'false hope' and 'pseudo-optimism', without collapsing into full 'despair', hopelessness, or fatalism, which lead to 'giving up' (Var der Lugt, 2021).[12] We expand on 'hybrid-valenced' accounts of climate hope in the next section.

In addition to fittingness, another well-known criterion for the evaluation of the cognitive rationality of emotions that is useful to understand our emotional reactions to warmer summers is warrant.[13] Rationality requires responding to apparent reasons and so an attitude is warranted when it is a response to apparent fit-related reasons. The distinction here is between a rational assessment ('warrant') and an assessment of objective normativity ('fit'). If facing danger makes fear fitting (i.e., if danger is a fit-related reason for fear), then fear is warranted for an agent if the agent is facing apparent danger (Naam, 2021).

What does the notion of warrant suggest regarding how to feel in the face of increasingly warmer summers? Fear could be understood as warranted in this context if its object, the exceptional warmth, manifests certain evidential cues of dangerousness. Most people are aware and hold both the belief that exceptionally higher temperatures are caused by climate change, and the belief that climate change is dangerous for life on Earth. But while current warmer summer temperatures are somewhat out of the ordinary, they have not yet become unbearable (at least not in Europe) and remain within the pleasurable-tolerable range for most.[14]

---

[12] Var der Lugt (2021) provides a thorough historical overview of the origins of philosophical pessimism. Drawing from the work of historical philosophical pessimists like Hume or Kant, she argues that in the 'dark times' of climate change and environmental degradation that haunt humanity, we should not think about optimism and pessimism as dichotomic notions, but rather as complementary.

[13] For the original distinction between *fittingness* and *warrant* see (D'Arms & Jacobson, 2000).

[14] This is compatible with there being many individuals who will consider current warmer days as unpleasant.

Additionally, survey studies report that about 60% of Americans prefer to live in a hotter climate, while only 29% would rather live in a colder one (Pew Research Center, 2009). So, although fear could be said to be a fitting response to this trend, a generalized strong preference for warmer weather among people, together with the experience of warmer, yet-not-unbearably-hot temperatures, could prevent agents from acknowledging that they are in a situation of apparent danger, preventing this emotional response from being warranted and thus also *required* by rationality.

Considerations of fittingness are important for tracking emotional rationality. However, in the practice of emotional evaluation, these considerations have traditionally been overshadowed by considerations of prudence or self-interest. Let's get back to Trump's and Putin's responses to Thunberg introduced in Section 2 above. Unlike Putin's, Trump's remarks did not suggest disproportionality or a mismatch between what Thunberg feels and aims at eliciting in others, and the object of those emotions. Thunberg was in fact cautious in making explicit reference to this proportionality: "I want you to feel the fear I feel every day, and then I want you to act (...) as if our house is on fire. *Because it is*".[15] However, Trump's words implied that Thunberg's anger was still inappropriate or unjustified and prompted her to 'chill'.

Demanding individuals to forgo their anger at an injustice for reasons of prudence or self-interest (e.g., due to anger being counterproductive in convincing people about the climate urgency) despite their anger being fitting or appropriate, is not an isolated practice. According to Amira Srinivasan, such practice belongs to the long philosophical and political tradition of *affective injustice*. Affective injustice is a second-order injustice parasitic on a first-order, conventional type of injustice emerging from the oppression of a victim. The wrongness of it lies in forcing people, through no fault of their own, into substantive and normatively costly conflicts—like the choice between self-preservation and justified rage (Srinivasan, 2018: 137). Given that, according to Srinivasan, apt anger has intrinsic value (due to e.g., it being a negative attitude towards something bad), those who, like Trump, demand angry individuals to forgo their fitting anger, face an argumentative burden: "they must explain why it is that in cases where one's anger would be counterproductive yet apt, prudential considerations must outweigh aptness considerations" (Srinivasan, 2018: 136). In absence of an account that explains the presupposed value superiority of prudence over fitting anger, Srinivasan argues, we can be suspicious that the counterproductivity argument against the expression of fitting anger masks an attempt of social control over certain socially excluded groups, traditionally slaves and women (Srinivasan, 2018: 136–144).

[15] Italics added for emphasis.

Cases of affective injustice are particularly interesting in this context. They involve a general conflict between consequentialist reasons (of different kinds) and non-consequentialist reasons for action generated out of apt or fitting emotions (anger or others) (Plunkett, 2020). In the case of climate change, anger may not be fitting were mitigation to be taken seriously and on time. However, even in this case, people would have a right to freely express or voice past injustices that have yet not been recognized. Deontological reasons could thus be added to this dichotomy, namely with the right to freedom of expression. For reasons of space, we deal with this dichotomy somewhere else.[16]

4. Psychological normativity and climate emotions

Psychological perspectives on climate emotions builds on work that has focused on identifying emotions, providing taxonomies, and describing correlations with other variables. In social sciences, there is a decent understanding of some climate emotions (mostly *anxiety* and *hope*) at the descriptive level and in relation to their contribution to certain outcomes, such as climate action or motivation. Climate change elicits numerous emotions among the public, and awareness and discussions on climate emotions are rising (Hyry, 2019; Ojala et al., 2022; Pihkala et al., 2022).[17] For example, the latest Youth Barometer in Finland found that the majority (59%) of the young had discussed climate anxiety over the past month (e.g., Pihkala et al., 2022).

In psychological research, emotions are usually understood as some type of "discrete, automatic responses to universally shared, culture-specific and individual-specific events" (Ekman & Cordaro, 2011). Emotions are evoked in response to real or imagined stimuli of relevance for us and they inform us about how to think and behave in different situations (e.g., what should be approached or avoided) (Damasio, 2005). The subjective experience of an emotion is called a feeling. Certain emotions, such as fear and happiness, are commonly considered as basic and adaptive and have survival value (Ekman & Cordaro, 2011). Thus, they are pre-programmed and involuntary, accompanied by corresponding physical reaction, although they can be modified through socialization and intentional efforts. Some emotions are fuzzy and less distinct, often involving several psychological and physical processes simultaneously. Anxiety is a paradigmatic example, which is understood as a mixture of affect (e.g., fear), cognition (e.g., worried thoughts), and physical changes (e.g., fast heartbeat).

[16] For an analysis of climate anger and affective injustice, see Authors (forthcoming).

[17] An extensive study into climate emotions in Finland (Hyry, 2021) found that, for example, 58 percent of the population expressed that they feel interest regarding climate change, and that feelings such as frustration (44%), powerlessness (39%), and hope (36%) are relatively widespread. Anxiety (25%) and shame (18%), commonly debated climate emotions were also expressed by many.

Claims about the normativity of emotions are usually implicit in psychology, but we have identified two types of normative assumptions. In the first type, emotional states *per se* are considered somehow inappropriate. They can deviate in statistical terms or in relation to an individual's previous emotional patterns, or be unfitting or irrational in the sense that they are not proportional in relation to their object in a similar way proposed by philosophers. For example, people can react to adverse events in seemingly irrational ways, such as with disproportional rumination or unfounded wishful and illusionary hope (but see Nolen-Hoeksema, Wisco, and Lyubomirsky 2008; Snyder et al. 2002 for thorough discussions). These criteria are particularly useful in clinical psychology (APA, 2013). [18] The second type of normative assumptions are not linked to the emotion per se, but to what their *consequences* are. For example, whether an emotion induces personal suffering or behaviours that cause harm to other people or society, or is considered unsuitable or immoral. In such cases efforts are typically made to change or manage it through medication, therapy, or some form of reinforcement.

Importantly, however, there are aspects of emotions and emotionality that may not always be meaningful to discuss in terms of normativity. If emotions are automatic reactions, it becomes clear that humans cannot fully and directly control them. Individuals also differ in their readiness to respond to stimuli. To exemplify, some are more regulated by moral views and the related feelings of shame and pride than others, and some tend to respond to crises by action while others ruminate. Contextual factors also influence; climate change, for example, is a complex chain of events caused by multiple practices across the world involving a certain degree of uncertainty and, hence, may induce different emotions to those produced by simpler and more manageable crises. Thus, certain normative judgements about emotions can be both unsuccessful and unjustified in psychological research. While rational criteria are useful in philosophy, practical and consequential approach are more commonly emphasized in psychology.

As to emotions that are felt in response to the *threat* of climate change, certain psychological criteria are commonly used and debated. Of these, we particularly focus on anxiety and hope. For example, climate worry has been discussed in pathological terms in society, with claims of "mass neurosis" or "hysteric bursts of emotion" (Pettersson et al., 2022; Verplanken & Roy, 2013). As described in the introduction, this can reflect an intentional mechanism to shift away the focus from the

[18] In certain mental illnesses, for example in depression and other mood disorders, individual's emotional patterns are intensified or reduced. Psychotic episodes and other altered states of consciousness can induce emotions that do not match the outer world. And finally, personality disorders can include emotional patterns that deviate from the normative expectations in society, such as is the case with the lower sense of remorse and empathy in antisocial personality disorder.

object of the emotional attitude to the emotional component of the message, and for claims of mental instability of the messenger. However, it is possible that some indeed consider climate anxiety as a sign of mental health problem. In our society, emotions and rationality have traditionally seen as alternating and separate, with the latter having a higher status. Also, in some public debates, anxiety is depicted as a feeling that people (particularly the young) should be protected from (Pihkala et al., 2020), such as was earlier discussed regarding the remarks of Putin in response to Thunberg's speech.

However, recent research has highlighted these views as simplistic. Emotions are increasingly seen as important sources of information and often rational responses to reality (Damasio, 2005; Verplanken & Roy, 2013). In fact, 93 percent of Europeans believe that climate change is a serious problem, and the government responses are widely understood to be inadequate (Special Eurobarometer, 2021, see also Hickman et al., 2022), which would seem to suggest climate-related anxiety as a rational response to climate change. Related to this, Bloodhart and colleagues (2019) found that messages framed with negative emotion matched better the participants' feelings about climate change, and conveyed impressions of the speaker as rational, strong, and caring. However, and in addition, researchers increasingly emphasize that anxiety can involve a variety of emotional and cognitive processes that range from minor and occasional states to more severe and chronic conditions (Sangervo et al., 2022). Thus, depending on the form climate anxiety takes, it can be considered a rational and a potentially adaptive response to a real crisis, or a threat to personal wellbeing and action (Clayton & Karazsia, 2020; Ojala et al., 2022; Pihkala, 2022; Wullenkord et al., 2021).

The way that people cope with their climate anxiety also matters. In this context, there has been debate about the appropriateness of climate hope. Although individuals need to be able to continue their lives despite the lingering threat, it is at least not *prima facie* obvious that hope is a fully appropriate emotional responses to the threat of climate change, as suggested in Section 3 above. On the basis of a psychological consequentialist approach, where emotions are evaluated in terms of their consequences, the concern has been that climate hope may hinder understanding of the gravity of the crisis and weaken motivation to engage in collective action and support the required social change (Hornsey & Fielding, 2016). Other recent research on climate hope argues hope to be an appropriate emotional reaction to climate change. The idea here is that hope is a highly complex emotion, the experience of which is a mixture of emotional, cognitive, existential, identity-related, and social aspects (Ojala, 2012b). According to this account, people may feel hope even in very serious and desperate circumstances – or perhaps precisely *because* they feel threatened. Without a risk of future harm, there would be no reason for

hope. Thus, while there might be a sense in which hope is irrational or non-fitting, this notion of hope might still be appropriate in some other sense, since hope can motivate efforts to improve the climate situation and can mitigate its seriousness (Geiger et al., 2021). This is consistent with findings showing that when people are more hopeful about how life will be like for future generations, they report more willingness to sacrifice for the sake of future generations (Fairbrother et al., 2021).

This research understands hope in a *general* sense (akin to a general state in which one is "to be hopeful", which relates to efficacy beliefs and trust). Indeed, Ojala (e.g., 2012; 2015) has found that individuals' engagement with climate action is more common when they feel both worry and '*constructive* hope'. Constructive hope entails coping through positive reappraisal/cognitive restructuring whereby the problem is acknowledged, but people can switch their perspective by also acknowledging some positive trends in mitigation and having trust in our collective ability and willingness to address the problem (Ojala, 2012). In other words, it is fruitful to assess the appropriateness of the coping strategies that are used to manage the crisis and negative emotions, for example experiencing hope in its different forms (Ojala, 2022). If negative emotions are managed by de-emphasizing the threat, this allows experiencing hope based on denial, but can lead to decreased engagement with the climate issue and does not seem to promote wellbeing either (Marlon et al., 2019; Ojala 2013). Importantly, from this viewpoint, constructive hope seems to be a highly appropriate climate emotion (judged based on both the match with reality and consequentialist criteria), at least if society and those in power indeed engage in climate mitigation.

Our emotional reactions can be also seen as a response directed towards our personal and collective role in contributing to climate change. These considerations are present in the debates, and it is relatively common to consider that public discussions blame ordinary people (Lehtonen et al., 2020). The shaming trends recently popularized in Swedish culture and media during the last couple of years, including the so-called "flight shame" (*flygskam*) (Wolrath Söderberg & Wormbs, 2019), are some instances of this. Some may think that it is fitting to feel bad because such feelings may be a necessary first step in environmental awareness. Shame and guilt are moral feelings and, while painful, the ability to experience them (at 'healthy' levels) promotes considerate relationships and enables a benevolent society by inhibiting maladaptive behaviours. But should we aim at deliberately attempting to make people feel guilt and shame when they contribute to high-emitting actions?

Although shame and guilt are emotions that are often used synonymously, there are some crucial differences between them (Tangney et al., 2007). Shame is a feeling that targets the whole "self" and creates a sense of worthlessness and helplessness,

which can induce unconstructive responses such as avoidance, blaming, and antagonism (Tangney et al., 1992). There is also no clear evidence suggesting that anticipation of shame would induce pro-environmental behaviour. Interviews of Swedish people who had stopped flying revealed that their decision was not based on shame, but rather on increased knowledge and insights about the climate crisis (Woltrath Söderberg & Wormbs, 2019). In line with these findings, climate-related flight shame seems to be rare and more consistently associated with personal norms (experienced moral obligation to avoid flying) than with social norms (perception that people think one ought to be ashamed or embarrassed about flying) (Doran et al., 2021).

Guilt, on the other hand, stems from the recognition that one has done something wrong and has been observed to influence motivation for action – including pro-environmental behaviour (Shipley et al., 2022). This highlights an important psychological aspect: people generally want to do the "right" thing and act prosocially. And when they do, they can feel pride and enjoyment that is sometimes called "warm glow" (Andreoni, 1990). In line with this, people feel good when acting on climate change, and anticipation of this feeling can motivate climate-friendly behaviour (Jia & van der Linden, 2020). At the same time, though, in our carbon-intense society individuals have only a few options to make a concrete impact on climate mitigation. Hence, they may not be able to change their behaviour as much as they feel they should.

Our collective and political failures to respond sufficiently to the climate threat are usually also object of emotional reaction. Related to this, the unequal distribution of risks and benefits can induce different emotions across society. Climate change is caused by wealthy nations and individuals, while other groups risk facing the most acute consequences (e.g., Althor et al., 2016; Schlosberg et al., 2014). Furthermore, the groups at risk do not have much influence: the disadvantaged people have rarely been heard in climate negotiations (Schlosberg, 2013), the future generations and non-human animals have no possibility of raising their voices and talking about their feelings, and the young have limited options to influence climate policy. These aspects can induce a variety of different emotions, of which we focus on anger.

When people face or detect injustices, this can trigger anger, which is also a highly activating feeling that can be needed to correct injustices (see, e.g., Stanley et al., 2021). Indeed, as the exchange described in Section 2 revealed, Thunberg expresses intense anger and disappointment due to the persistent delay in climate change mitigation. Her appeal to an increase in the recognition of the threat and injustice behind climate change is a common feature of the angry political rhetoric employed by historical figures like Baldwin, Malcom X, Catherine MacKinnon and Angela

Davis, whose anger represents a verbal expression, "a swift and often automatic conversion of sentiment into word" (Srinivasan, 2018: 140). Both Trump and Putin would have had the power to speed up mitigation, but instead of expressing an intention to do this, they dismissed Thunberg's anger and its causes. The inadequacy in societal responses has made some people (particularly among the young) to feel powerlessness, frustration, and even that they are being betrayed (Hickman et al., 2021). However, to fully understand different reasons to feel threatened, perhaps it should be acknowledged that some are *personally* more harmed by solutions than climate change. For example, those with high-emitting jobs may risk losing their jobs, economic investments in the fossil-fuel industry will decline, and those who have invested in denying the climate threat risk losing their credibility and social status. Because of this, Thunberg's and other climate advocates' claims can be perceived as a threat or even as overt aggression by some.

5. Conclusion

The landscape of climate emotions is broad and complex, with different emotions being elicited in different people to different degrees. This gives rise to an additional, yet undertheorized layer of disagreement among the public: there is division of opinion not only about what should be believed regarding climate change, but also about how we should emotionally react to it.

In this paper, we have tried to clarify this disagreement by examining the normativity of climate emotions and the different reasons in favour of different affective reactions to climate change. For that, we provided an overview of the main normative criteria for the evaluation of emotions existing in the philosophical and the psychological literature, including *fittingness*, *warrant*, and considerations of *prudence* from philosophy, and *clinical*, *consequentialist*, and *contextual* considerations from psychology. We used these criteria to evaluate some paradigmatic climate emotions, including climate-related *hope*, *anxiety*, *shame*, and in climate emotional dilemmas, like how to feel in the face of increasingly warmer summers. We showed that different normative criteria for the evaluation of these emotions can yield different answers to the question of what is appropriate to feel in the face of climate change, partly depending on how the object of that emotion is specified or individuated (e.g., warmer temperatures alone, or climate change as a whole).

Importantly, we conceived these normative criteria as providing *pro-tanto* reasons for the appropriateness or inappropriateness of certain climate emotions, and thus conclusions following from these criteria ought not be regarded as *all-things-considered* judgements about the 'validity' of these emotions. The appropriateness, relevance, or salience of these normative criteria for the evaluation of emotions will

be determined, among other things, by the social norms operating in different social contexts. If we are to engage in the project of eliciting different climate emotions (in which we've shown we are already immersed), it is important to note that emotions do not exist isolated from the individuals who experience them. Thus, the contributing value of a given emotion will vary depending on, for example, personality dispositions and other simultaneous emotions the individual may hold. So, there should be room for including considerations about individual differences and limitations in emotional responsivity to the normative assessment. The same way as people may engage in the type of climate action that suits them best, different emotional states may serve some people better than others. Additionally, at the population level, we may not need everyone across the board to feel equally hopeful or equally angry. We can also speculate that some emotions are more natural and beneficial to certain societal actors than to others, like perhaps 'hopeful pessimism' to philosophers, and 'constructive hope' to educators. More interdisciplinary research can help to provide evidence-based guide on the appropriateness of different emotions at the population level.

In sum, the aim of the normative exercise developed in this paper was to provide tools for the better understanding of one's and others' climate emotions. Emotional disagreements as the ones becoming more predominant in the public debate and the eventual judgment (public or internal) or sanctions that may follow from this disagreement, have potential consequences at the individual and social level. Negative feelings arising from the inappropriateness of one's emotions may hinder motivation for action and negatively impact self-esteem. Emotional disagreement, judgment, and eventual sanctioning (in the form of public shaming, gossiping, ostracism, etc.) can potentially hinder our already fragile trust that others be appropriately motivated to act. This is problematic, given the role of trust in cooperation and collective action, much needed to solve the climate crisis. It is possible to alleviate some of these consequences, we argue, if we come to terms with the complex nature of climate emotions and their normative justification.

References

Althor, G., Watson, J.E.M., & Fuller, R. A. (2016). Global mismatch between greenhouse gas emissions and the burden of climate change, *Scientific Reports*, 6, 20281.

American Psychiatric Association (2013). *The Diagnostic and Statistical Manual of Mental Disorders,* Fifth Edition (DSM-5).

Archer, A. & Mills, G. (2019). "Anger, Affective Injustice, and Emotion Regulation", *Philosophical Topics*, vol 47(2), 75–94.

Bergmann, Z. & Ossewaarde, R. (2020). Youth climate activists meet environmental governance: ageist depictions of the FFF movement and Greta Thunberg in German newspaper coverage, *Journal of Multicultural Discourses*.

Bloodhart, B., Swim, J., & Dicicco, E. (2019). "Be Worried, be VERY Worried:" Preferences for and Impacts of Negative Emotional Climate Change Communication. *Frontiers in Communication*, 3: 63.

Bykvist, K. (2009). "No Good Fit: Why the Fitting Attitude Analysis of Value Fails" *Mind*, Vol. 118, No. 469, pp. 1–30.

Campbell, T. H., & Kay, A. C. (2014). Solution aversion: On the relation between ideology and motivated disbelief, *Journal of Personality and Social Psychology*, 107(5), 809–824.

Clayton, S.D., & Karazsia, B.T. (2020). Development and validation of a measure of climate change anxiety. *Journal of Environmental Psychology*, 69: 101434.

Cullity, G. (2018). *Concern, Respect, and Cooperation*, Oxford: Oxford University Press.

Damasio, A. (2005a). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York, NY: Penguin.

D'Arms, J. and Jacobson, D. (2000). "The Moralistic Fallacy: On the 'Appropriateness' of Emotions", *Philosophy and Phenomenological Research*, LXI (1): 65–90.

Doran R, Pallesen S, Böhm G, Ogunbode CA. (2021). When and why do people experience flight shame? *Annals of Tourism Research*, 92, 103254.

Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic? *Emotion Review*, 3(4), 364–370.

Gallegos, F. (2021) "Affective injustice and fundamental affective goods", *Journal of Social Philosophy*, 00, 1–17.

Geiger, N. Swim, J.K., Gasper, K., Fraser, J., & Flinner, K. (2021). How do I feel when I think about taking action? Hope and boredom, not anxiety and helplessness, predict intentions to take climate action, *Journal of Environmental Psychology*, 76, 101649.

Glick, P., & Fiske, S. T. (2001). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality, *American Psychologist*, 56(2), 109.

Greenspan, P. (1998). *Emotions and Reason: An Inquiry into Emotional Justification*, London: Routledge & Kegan Paul.

Gregersen, T., Doran, R., Böhm, G., Tvinnereim, E., & Poortinga, W. (2020). Political orientation moderates the relationship between climate change beliefs and worry about climate change, *Frontiers in Psychology*, 11:1573.

Fairbrother, M., Arrhenius, G., Bykvist, K., & Campbell, T. (2021). Governing for future generations: How political trust shapes attitudes towards climate and debt policies, *Frontiers in Political Science* 3.

Hickman, C., Marks, E., Pihkala, P., Clayton, S., Lewandowski, R.E., Mayall, E.E., Wray, B., Mellor, C., van Susteren, L. (2021). Climate anxiety in children and young people and their beliefs about government responses to climate change: A global survey. *The Lancet. Planetary Health* 2021, 5, e863-e873.

Hornsey, M.J., Fielding, K.S. (2016). A cautionary note about messages of hope: focusing on progress in reducing carbon emissions weakens mitigation motivation. *Global Environmental Change*, 39:26–34.

Hyry, J. (2021). *Climate emotions: Summary of key findings*. Retrieved 10 April 2022, from: https://media.sitra.fi/2019/11/29131052/sitraclimate-emotions-report-2019.pdf

Jacobson, D. (1997). "In Praise of Immoral Art", *Philosophical Topics*, 25, pp. 155–99.

Kunelius, R., & Roosvall, A. (2021). Media and the Climate Crisis, *Nordic Journal of Media Studies*.

Lehtonen, T. Niemi, M.K. Perälä, A. Pitkänen, V. & Westinen, J. (2020). Making Sense of Climate Change: Perspectives of Citizens, *Business Leaders and Political Decision-Makers*. Vaasa: e2 Research and University of Vaasa.

McDowell, J. (1978) "Are Moral Requirements Hypothetical Imperatives?" *Proceedings of the Aristotelian Society*, supp. vol. 52, pp. 13–29.

Naar, H. (2021) "The fittingness of emotions", *Synthese*, 99 (5-6): 13601–13619.

Neckel, S. and Hasenfratz, M. (2021). "Climate emotions and emotional climates: The emotional map of ecological crises and the blind spots on our sociological landscapes", *Social Science Information*, 60(2), 253–271.

Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirsky, S. (2008). Rethinking Rumination, *Perspectives on Psychological Science*, 3(5), 400–424.

Ojala, M. (2012a). Regulating worry, promoting hope: How do children, adolescents, and young adults cope with climate change? *International Journal of Environmental & Science Education*, 7(4), 537–561.

Ojala, M. (2012b). Hope and climate change: the importance of hope for environmental engagement among young people, *Environmental Education Research*, 18(5), 625–642.

Ojala, M. (2013). Coping with climate change among adolescents: Implications for subjective well-being and environmental engagement, *Sustainability*, 5, 2191–2209.

Ojala, M., Cunsolo, A., Ogunbode, C.A., & Middleton, J. (2022). Anxiety, worry, and grief in a time of environmental and climate crisis: A narrative review, *Annual Review of Environment and Resources*. 46:10.1–10.24.

Painter, J., Ettinger, J., Doutreix, M.-N., Strauß, N., Wonneberger, A., & Walton, P. (2021). Is it climate change? Coverage by online news sites of the 2019 European summer heatwaves in France, Germany, the Netherlands, and the UK, *Climatic Change*, 169: 4.

Pettersson, K., Martikainen, J., Hakoköngäs, E., & Sakki, I. (2022). Female politicians as climate fools: Intertextual and multimodal constructions of misogyny disguised as humor in political communication, *Political Psychology*.

Pew Research Center (2009). *Most like it hot*. https://www.pewresearch.org/social-trends/2009/03/18/most-like-it-hot/

Pihkala, P., Sangervo, J., & Jylhä, K.M. (2022). Nuorten ilmastoahdistus ja ympäristötunteet. In T. Kiilakoski (Ed.), *Youth Barometer 2022*. Helsinki: Ministry of Education and Culture, The State Youth Council, The Finnish Youth Research Network.

Pihkala, P., Cantell, H., Jylhä, K.M., Lyytimäki, J., Paloniemi, R., Pulkka, A., & Ratinen, I. (2020). Ahdistuksen vai innostuksen ilmasto? Ilmastoviestinnän ja -kasvatuksen keinoja ilmastoahdistuksesta selviytymiseen. In E. Pekkarinen, & T. Tuukkanen (Eds.), *Future of the planet and the rights of the child*, Office of the Ombudsman for Children in Finland 2020:4, Jyväskylä, Finland.

Plant, E. A., Hyde, J. S., Kelner, D., & Devine, P. G., (2000). The gender stereotyping of emotions, *Psychology of Women Quarterly*, 24, 81–92.

Plunkett, D. (2020) "Debate: Anger, Fitting Attitudes, and Srinivasan's Category of "Affective Injustice", *The Journal of Political Philosophy*, 29(1) pp. 117–131.

Putin, V. (2020). Intervention at the workshop 'Energy Partnership for Sustainable Growth', Energy Forum of Moscow, Moscow.

Roberts, R. (1988) "What An Emotion Is: A Sketch," *The Philosophical Review*, pp. 183–209.

Sangervo, J., Jylhä, K.M. Pihkala, P. (2022). Climate anxiety: Conceptual considerations, and connections with climate hope and action, *Global Environmental Change*, in press.

Shipley, N. & van Riper, C.J (2022). Pride and guilt predict pro-environmental behavior: A meta-analysis of correlational and experimental evidence, *Journal of Environmental Psychology*, 79, 101753.

Schlosberg, D. (2013). Theorising environmental justice: The expanding sphere of a Discourse, *Environmental Politics*, 22, 37–55.

Schlosberg, D., & Collins, L.B. (2014). From environmental to climate justice: Climate change and the discourse of environmental justice. *WIREs Climate Change*, 5, 359–374.

Snyder, C.R., K.L. Rand, E.A. King, D.B. Feldman, & Woodward, J.T. (2002). 'False' hope, *Journal of Clinical Psychology*, 58(9), 1003–1022.

Stanley, S., Hogg, T.L., Levinston, Z., & Walker, I. (2021). From anger to action: Differential impacts of eco-anxiety, eco-depression, and eco-anger on climate action and wellbeing, *The Journal of Climate Change and Health*, 1, 100003.

Srinivasan, A. (2018). "The Aptness of Anger", *The Journal of Political Philosophy*: Volume 26, Number 2, 2018, pp. 123–144.

Solomon, Robert C., (1976). *The Passions, Garden City*, New York: Doubleday Anchor.

Tangney, J. P., Wagner, P., Fletcher, R., & Gramzow, R. (1992). Shamed into anger? The relation of shame and guilt to anger and self-reported aggression, *Journal of Personality and Social Psychology*, 62(4),669–75. doi: 10.1037//0022-3514.62.4.669.

Tangney, J. P., Stuewig, J., Mashek, D. J. (2007). Moral emotions and moral behavior, *Annual Review of Psychology*, 58, 345–372.

Toivonen, H. (2022). Themes of climate change agency: a qualitative study on how people construct agency in relation to climate change, *Humanities & Social Sciences Communications*, 9: 102.

UNDP (2022). "New threats to human security in the Anthropocene: Demanding greater solidarity", *Special Report*, New York: United Nations Development Programme.

Var der Lugt, M. (2021). *Dark Matters: Pessimism and the Problem of Suffering*, Princeton University Press.

Verplanken, B., & Roy, D. (2013). 'My worries are rational, climate change is not': Habitual ecological worrying is an adaptive response. *PLoS ONE*, 8(9), 1–6.

Wilcke, R.A.I., Kjellström, E., Lin, C., Matei, D., & Moberg, A. (2020). The extremely warm summer of 2018 in Sweden – set in a historical context. Earth System Dynamics, 11, 1107–1121.

Wullenkord, M., Tröger, J., Hamann, K. R., Loy, L., & Reese, G. (2021). Anxiety and Climate change: A validation of the climate anxiety scale in a German-speaking quota sample and an investigation of psychological correlates. *Climatic Change*, 168, 20.

Wolrath Söderberg, M., & Wormbs, N. (2019). Grounded: Beyond flygskam. *European Liberal Forum & Fores*.

John Broome[1]

How to Value a Person's Life[2]

The work of economists on the value of human life divides into two strands. The first values life on the basis of people's willingness to pay to reduce risk to their lives, and it aims to derive a value that is suitable for use in cost-benefit analysis. The second values life on the basis of the length of the life and a measure of its quality. It is mainly used for cost-effectiveness analysis in health care. This paper condemns the theoretical underpinning of the first strand. Nevertheless, it develops a reconciliation of the two strands. It shows how the idea of valuing life by willingness to pay can be reconciled with the thinking that underlies the second strand. The result is a method of valuing life that has a defensible foundation and could be implemented in practical cost-benefit analysis.

[1] john.broome@philosophy.ox.ac.uk

[2] The 2022 Brocher Lecture.

1. Introduction

Economists began making calculations involving the value of people's lives in the 1960s and 1970s. There have always been two strands of thinking. On the one hand there were transport economists, environmental economists and others who started incorporating the value of lives into cost-benefit analyses on the basis of people's willingness to pay for extending their lives.[3] On the other hand there were health economists who developed measures – principally 'quality adjusted life years', or 'qalys' – for the benefits of health care to use in cost-effectiveness analysis of different treatments.[4]

Recently there have been some signs of convergence between the strands, and I hope to make a small contribution to their reconciliation. This is particularly worth doing because the economic value of life has acquired much greater importance in recent decades. For one thing, it is a major component of the social cost of carbon, which is the key figure in climate change economics. And climate change is the leading problem of our age.

2. Willingness to pay

The first strand of thinking is embodied in the notion of 'the value of a statistical life'. Like many other people, I hate that term. But the part of it I hate is different from the part many other people hate. Many of them hate the word 'life', because they don't like to be seen as setting a value on people's lives. They prefer to set a value only on a risk of losing one's life, rather than on losing a life itself.

But I hate the word 'statistical'. It reminds me irresistibly of Joseph Stalin's famous (apocryphal) remark:

A single death is a tragedy; a million deaths is a statistic.

Contrary to what Stalin implies, a million deaths is a million tragedies. The badness of deaths is proportional to the numbers of deaths.

And that is true of risks too. The badness of a 1 in 10,000 risk of losing one's life is just 1/10,000 of the badness of losing one's life. This is an elementary consequence of expected utility theory. Why is a risk bad? Because of the badness of whatever it is a risk of. The primary object of value is what may happen – the possible outcome of the risk; the value of the risk derives from the value of the outcome. The nature of the derivation is easy to state: the badness (negative value) of a risk is the badness of

[3] The first example of this strand that I know is 'L'utilité sociale d'une vie humaine' by Jacques Drèze (1962).

[4] The history of the development of qalys is described in detail by Eleanor MacKillop and Sally Sheard in 'Quantifying life' (2018).

the bad thing it is a risk of, multiplied by its probability. That is to say, the badness of a risk of death is proportional to the probability of death.

For instance, it is an implication of expected utility theory that the badness of 10,000 people's being exposed to a 1/10,000 risk of dying is the same as the badness of one person's being exposed to a certainty of death (if all the people are similar). I shall later qualify this conclusion on grounds of fairness. But for the time being I shall suspend the qualification and stick with this basic conclusion of expected utility theory.

It conflicts with traditional cost-benefit analysis. Traditionally, cost-benefit analysis values a benefit to a person by how much money the person would be willing to pay for it, and it values a harm to her by how much she would accept as compensation for bearing it. These amounts are technically the compensating variation (CV) of the benefit or minus the compensation variation of harm. Traditionally, cost-benefit analysis reckons a change, which brings benefit to some and costs to others, as a good thing if and only if the sum of all the people CVs is positive. The sum of CVs is the criterion for accepting a project.

The CV of a risk of death is not proportional to the probability of death. Valuation by the CV is therefore not consistent with expected utility theory.

It's easy to see why the CV is not proportional to probability. It is an implication of expected utility theory itself, applied to the person's own decision making. You can do the algebra, but the reason is easy to see without the algebra. Imagine you have to compensate someone for bearing a risk of death, and think how much the compensation she would require will increase as the risk gets higher. If she dies, she will get much less benefit from the compensation than if she lives, because she won't get to spend it. Indeed, money may be worthless to her if she dies. As the risk gets higher, the expected benefit she receives from any particular amount of compensation therefore gets less and less. So she will require proportionally more compensation to make up for the chance of getting less benefit. In the extreme, if money is totally worthless to her if she dies, it will be completely impossible to compensate her with money for a very high risk of dying.

To put it briefly, the value of money to the person diminishes as the probability of her dying increases. It gets progressively harder to compensate her because you are trying to do so using a medium that has progressively less value to her. This makes it obvious that the CV is not a good measure of the value of risk of dying. If you value risk of dying using the 'measuring-rod of money' as A. C. Pigou calls it, you will find your measuring-rod constantly varying in length. So it will not work properly. The value of risk is proportional to the probability. Measuring value by the CV implies it is not, but this only shows it is an unsatisfactory means of measurement.

Oddly, though, this method of measurement is historically what led economists to concentrate their attention on the value of a *statistical* life. It happened around 1970. Lots of projects cause deaths. For example, big engineering projects very often lead to deaths in the course of their construction. These deaths are a cost of the project. If you value them by their CVs, you get very big numbers because it takes a very large – probably infinite – amount of money to compensate someone for dying. So if you think the CV is a correct means of valuation, you will think you have to reject any project that causes a death. But that is clearly not so. Some projects that cause deaths are worthwhile nonetheless. So what do you do?

What you ought to do is realize that the CV is not a good measure of value. But there is a strong ideology behind traditional cost-benefit analysis, so that is not what economists did. Instead they decided to measure the value of the risk of death rather than the value of deaths themselves. This cut out those very high valuations and allows some projects to be accepted even if they cause deaths.

This move to valuing statistical rather than individual lives was made to preserve the ideology. It wasn't worth preserving. Traditional cost-benefit analysis based on the sum of CVs should have been abandoned long ago, for many reasons. I've said it is inconsistent with expected utility theory. This is a bad fault, but it pales into insignificance compared with some of its other faults. As long ago as 1941, Tibor Scitovsky showed that it leads to flatly contradictory results.[5] The sum of CVs in moving from some situation B to another A may be positive, and at the same time so may the sum of CVs in moving from A to B. The sum of CVs criterion therefore implies that A is better than B and also that B is better than A. This is a reductio ad absurdum of the criterion.

This particular absurdity can be circumvented by applying a double – backward and forward – criterion. One option A is declared better than another option B if the sum of CVs in moving from B to A is positive and the sum of CVs in moving from A to B is not positive. But in 1955, Terence Gorman showed that this double test can imply a different sort of inconsistency.[6] It can imply that A is better than B, B better than C and C better than A. This criterion too is shown to be false by reductio ad absurdum.

That was almost 70 years ago and it should have finished off traditional cost-benefit analysis. In any case, there was never a good argument for using the sum of CVs as a criterion. The sum of CVs is generally taken to be a test of whether the gainers from a project could compensate the losers, so that nobody ends up worse off. For this reason it is often called 'the compensation test' or the test of a 'potential

[5] Scitovsky (1941), 'A note on welfare propositions in economics'.

[6] Gorman (1955), 'The intransitivity of certain criteria'.

Pareto improvement'. But actually it is not. Even if the sum of CVs is positive, the gainers may not be able to compensate the losers. This may surprise you at first, but remember that a transfer from the gainers to the losers changes the distribution of wealth and consequently it changes market prices. The sum of CVs is calculated at the prices that prevail before the transfer. If the gainers tried to make the transfer, prices would change and it might not be possible to end up with a Pareto improvement. This was demonstrated by Robin Boadway in 1974.[7] So the sum of CVs is not the compensation test.

Besides, there was never any good reason for accepting the compensation test in the first place. It was recommended by Nicholas Kaldor[8] and supported by John Hicks[9] in 1939, but neither of them offered any real argument for it. It is easy to produce counterexamples in which the gainers from a change could compensate the losers, but nevertheless the change is obviously not for the better.

In sum, the basis for traditional cost-benefit analysis – using the sum of CVs as a criterion – was thoroughly discredited decades ago. Traditional cost-benefit analysis should be abolished. The sum of CVs had ideological support because it purported to value a project without the need for interpersonal comparisons of wellbeing, even when the project is good for some people and bad for others. That ambition has to be abandoned. It was hopeless from the start. When a project is good for some and bad for others, obviously we have to compare the good of some with the bad of others. That is exactly what we are doing when we evaluate the project.

This does not imply that the CV of a risk to life is useless for the purpose of valuation. It provides useful information about the value a person sets on her life. It does not *determine* the value of her life, but it can be good evidence about the value of her life.

If we are to use CVs (willingnesses to pay) as evidence, we have to be ready to make adjustments according to the value of money to different people. It is widely recognized that the value of money to rich people is less than to poor people because the rich already have a lot of the things money can buy. Also, I have just explained another source of variation in the value of money. It depends on how near death a person is: if she is old or for some other reason exposed to a bigger risk of dying soon, money is worth less to her because she is less likely to have time to spend it. People's CVs must be adjusted according to these differences.

For instance, it is obvious that the CV of reducing risk will on average be lower in a poor country than a rich one. For this reason, in an international project, saving life in a poor country will get lower priority than saving life in a rich one if we apply

[7] Boadway (1974), 'The welfare foundations of cost benefit analysis'.

[8] Kaldor (1939), 'Welfare propositions of economics'.

[9] Hicks (1939), 'The foundations of welfare economics'.

the criterion of the sum of CVs. But it's equally obvious that the reason the CV is lower in a poor country is that on average money is more valuable to the people there. So if we apply an appropriate adjustment, the same priority will not emerge.

3. Fairness

Now back to a point I made earlier and immediately suppressed. I said it was an implication of expected utility theory that the badness of 10,000 people's being exposed to a 1/10,000 risk of dying is the same as the badness of one person's being exposed to a certainty of death. There is an argument to say that expected utility theory goes wrong here, because it is better for the risk of death to be widely distributed rather than focussed on one person. This is on ground of fairness. That is exactly the argument Peter Diamond used against expected utility theory in 1967,[10] and it is a good one.

It even supplies a sort of backhanded support to using the sum of CVs. Just because the unadjusted CV of risk increases more than in proportion to the risk, it will reckon a more concentrated risk as worse than a more widely distributed one. Indeed, in 1982 the UK National Radiological Protection Board (NRPB) reached exactly this conclusion by this method.[11] It was assessing the badness of radiation leaking from nuclear plants, using valuation by CV. If there were to be a particular number of deaths, it much preferred them to be widely distributed across the UK population, rather than concentrated on the close neighbours of the plant. This conclusion aligns with what fairness might also recommend.

But this support for unadjusted CV valuation is ineffective. The CV of a risk has nothing to do with fairness. Fairness and the sum of CVs may reach the same conclusion, but that is just a coincidence. The NRPB's reason for preferring a wide distribution of deaths is specious.

Still, what we should do about fairness is a real question. I think Diamond was wrong to see it as an objection to expected utility theory. I think the best practical way of taking fairness on board is to treat it separately from goodness (or value). Of course, fairness is good, but it behaves in a very different way from other goods, so it is advantageous to separate it. Expected utility theory applies to good, and fairness has to be taken into account separately. To take a much-discussed example. Suppose a health service has life-saving treatment available, but not enough to treat everyone who needs it. If it is decided on grounds of qalys whom to give it to, it will go to people who are otherwise in good health rather than those who have other health problems.

[10] Diamond (1967), 'Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility'.

[11] Fleishman and Clark (1982), 'Evaluating future detriment from radioactive discharges'.

But it is unfair to deny life-saving treatment to some people just because their lives are already less good than they might be. However, we should not allow this consideration of fairness to distort our judgement about the benefits of the treatment. It does indeed do more good to treat people who are otherwise healthy than those who are not. But good is not everything; fairness also matters. Sometimes we ought not to do the best thing, because doing so is too unfair. That seems to be so in this particular case I described: on grounds of fairness, we should not discriminate against the less healthy candidates for treatment.

The upshot of this is that fairness does not constitute an objection to what I have said about the sum of CVs as a criterion for cost-benefit analysis. My argument was concerned with value or goodness, but we have to remember that fairness is a further consideration that needs to be taken into account. I now once more revert to goodness.

4. The value of a person's life

Now I come to the alternative strand of thinking. Here I shall often speak in my own voice, since at a general level I subscribe to this strand.

Here are some features of this alternative approach. First, it is more interventionist. It involves thinking about how good lives actually are, recognizing that people may make mistakes about the goodness of their own lives. Economists like to base their valuations on people's preferences alone. But in thinking about the value of lives, we have to remember that people have different preferences at different stages of their lives. At the very least, those preferences will have to be integrated together, and this integration cannot itself be based on the preferences. So some intervention is inevitable.

A second feature of the alternative approach is that risk is not essential to the method of valuation. Of course, risk and uncertainty always has to be taken into account, but it is taken into account in a more conventional way. When there is risk, various outcomes are possible. Each of these has a value and a probability, and together their values and probabilities determine the value of the risk in the way described by expected utility theory. The primary bearers of value are outcomes, which are states of affairs that themselves have no risk in them. So the first job for this approach is to work out the value of life in a state of affairs without risk or uncertainty. Any practical decisions will require risk to be accounted for later.

Let me illustrate the general problem of valuation as I see it. Even to illustrate it, I need to make an assumption. I shall assume that the value of the world depends only on each person's temporal wellbeing at each time she is alive, by which I mean how well the person's life is going at each time.

This assumption sets aside the wellbeing of animals, the intrinsic value of nature, and the value of human cultures, in so far as they have values beyond their effects on human wellbeing. It even sets aside the value that the human species may have, apart from the wellbeing of the individuals who make it up. So it may not capture all the badness of human extinction. I am not denying the existence of all these values, but I am assuming that they are separable from the value of human wellbeing, so they can be taken account of separately. I've chosen to concentrate on the value of human wellbeing.

I also mean to allow for different conceptions of wellbeing, from a hedonist conception consisting of pleasure less pain, to a very general conception that includes health, access to travel and social life, having as nice house, and so on. Given all these caveats, I think the assumption that the value of a state of affairs depends only on people's temporal wellbeings is fairly uncontentious.

If you grant it, I can illustrate the general form of our problem with a picture.[12] This picture is supposed to illustrate the problem of climate change. It shows two possible states of affairs. Time is measured horizontally, with the vertical line marking the present. Each horizontal line belongs to a person, and the graph sitting on that line shows the person's temporal wellbeing through her life. There are some presently-existing people and some future people. Some people exist in one possible world and not in the other. In the world of *business as usual*, the quality of life in the future is less good than in *respond*, and lives are shorter.

Next I assume separability of people. That is, I shall assume that the goodness of the world is made up of the goodness of each of the people's lives. Again, this is not a very contentious assumption. It doesn't rule out causal interactions between people. Each person's temporal wellbeings may well be affected by how other people's lives are going; that is not excluded. But once we have identified the temporal wellbeings, we can evaluate the state of affairs person by person. This means we can split our task into two stages. First, work out how good is each person's life. Then work out how the goodnesses of all the people's lives go together to determine the goodness of the state of affairs.

This second stage is the business of a social welfare function. For instance, we might want an egalitarian formula, or one that does not favour equality, such as a utilitarian formula. But that's just the beginning; we then have to take account of changes in population. Some people adopt average utilitarianism, for example; others total utilitarianism, in which case they have to settle on a zero of wellbeing. All this is very difficult, but fortunately I'm not concerned with the social welfare function. I am concerned with the first stage, to work out the goodness of the individual lives.

[12] It is taken from my *Weighing Lives*, p. 10.

So what is the value of a person's life? I have already assumed implicitly that it depends on how the life goes, which is to say the person's temporal wellbeings at each time. I said that was fairly uncontentious, but the contention starts when we come to working out the form of the function from temporal wellbeings to the overall value to a person's life. How do temporal wellbeings aggregate? The simplest function is just additive: the value of a life is the arithmetic total of its temporal wellbeings. We may call this 'intrapersonal utilitarianism'. But lots of other functions are possible, which take account of the shape of the life. For example, it may be better for life to get progressively better rather than progressively worse. It may be that how life ends is particularly important in determining how good the life is as a whole. Alternatively, the beginning may be the most important, and later times of life may be less so. It may be good to have a high peak, or alternatively it may be good for a life to maintain an even tenor. And so on.

I have to confess that I know of few theoretical arguments that adjudicate among all these possibilities. Mostly it seems down to intuition to settle on the correct formula. I do think there are good arguments for interpersonal utilitarianism, which means that social value is the sum of individual wellbeings. But similar arguments for intrapersonal utilitarianism are much less convincing.

Nevertheless, I suggest we adopt intrapersonal utilitarianism as a default theory, in the absence of an argument to show it's wrong. I don't insist it is correct. I only suggest we need some good reason if we are to depart from it.

If it is right, then the value of extending a person's life is simply the total wellbeing she acquires during her extra period of life. Putting it another way, it's the total of wellbeing-adjusted life years in that period: 'walys' as I playfully called them in my book *Weighing Lives*. I suggest this as a default because it seems the most conservative, neutral formula. It is also intuitively attractive. What could be more natural than to think the goodness of a life is its total goodness, integrated over time? Clearly many people working on the value of life have taken it for granted. For decades public health economists and others have taken for granted more specialized versions of it, in the form of qalys and dalys (disability-adjusted life years). Although there has always been debate about the right way to make the quality adjustment in qalys, there's been little disagreement about the use of years.

5. Interpersonal comparisons

So the value of a life can be described as the total of wellbeing-adjusted life-years. For practical decision-making, of course, the difficult bit is the wellbeing adjustment. But even before we get to that, there is an important practical implication of valuing lives on the basis of life-years. Even traditional cost-benefit analysis is

equipped to work with life-years instead of undifferentiated life-saving. The notion of the VSLY – the value of a statistical life-year – is well recognized. It would surely be better to work with VSLY rather than VSL – the value of a statistical life. Even if intrapersonal utilitarianism is not correct, saving a life is much less valuable if it extends the life by just a few weeks than if it extends it by many decades. This is a minimal improvement to practice that could easily be adopted. Some practitioners seem nevertheless reluctant to adopt it. Perhaps this is because it favours saving the lives of young people over the lives of old people, and empirically the old may be just as willing as the young to pay to reduce risk to their lives. But if that is so, it is because old people have more money and less to do with their money. It is because money has less value to them, not because their lives have the same value to them as young people's lives have to them.

Now, what about a practically implementable measure of temporal wellbeings? I am not going to give a definitive answer to this question. I could not give one in any case because it plainly depends on what a person's temporal wellbeing consists in. There are many theories about this, which have been much debated. Any answer to the question of measuring wellbeing has to be tied to a particular theory of what wellbeing is.

To narrow the task, I shall look for a reconciliation with the first strand of thinking about the value of life. I want to stay as close as possible to the conventional methods of cost-benefit analysis, because this will make my suggestion easiest to implement in practice.

An underlying principle of conventional methods is that prices and compensating variations in general can be understood as measures of wellbeing. The price a person is willing to pay for some good is supposed to measure the marginal contribution that good makes to her wellbeing. More exactly, the prices of goods are proportional to the relative contributions the goods make to each person's wellbeing. This assumes that the person's wellbeing is aligned with the preferences that underlie her choices. This might be either because her wellbeing actually consists in the satisfaction of her preferences, or that her preferences are accurately formed on the basis of her wellbeing. If we are to be reconciled with conventional cost-benefit analysis, we shall have to accept this assumption.

Sticking to conventional methods consequently puts a demand on our measure of the value of life. It has to be commensurate with the prices of the ordinary goods that figure in the costs and benefits included in cost-benefit analysis. Our measure of temporal wellbeing will have to be commensurate with ordinary goods, in such a way that the prices of goods are proportional to their marginal contribution to wellbeing.

To see the point of this requirement, notice that the equivalent income measure

of wellbeing does not meet it. Equivalent income measures wellbeing as a quantity of money,[13] which is a good start, but the prices of ordinary goods are not proportional to their contribution to wellbeing measured this way.

We can satisfy this requirement by building a measure of a person's temporal wellbeing on the basis of the person's willingness to pay to extend her life at that level of wellbeing. That is my suggestion.

We shall need willingnesses to pay for each different type of life the person might lead during the extended period. By a 'type of life', I mean the set of all those natural features of a period of life that contribute to determining the goodness of the person's life – which is to say her temporal wellbeing – during the period. Each type of life is assigned value by means of the person's willingness to pay for an extension to her life of this type. These willingnesses to pay must be discovered empirically. For instance, people might be asked what they would be willing to pay to extend their life by one year, living a life of such-and-such a type. (The extension need not be at the end. It might be inserted in the middle of a life.)

These willingnesses-to-pay will give a value to each type of life, in terms on money. These values will be measured on a ratio scale. The zero of the scale is given by the life's not being extended at all. This ratio scale is particular to each person, and it will assign a value to each type of life the person might lead.

We need different people's scales to be comparable. So we next need to bring different people's scales into line. The zero of the scale is assigned to life's coming to an end, which is equally bad for everyone. So this zero level is already interpersonally comparable. Consequently, it is only the size of the unit of value that needs to be aligned between people.

I assume that leading a particular type of life is equally as good for one person as it would be for anyone else. If there is a type of life that is possible for everyone, this type will have a place in everyone's scale of value. We have only to adjust each person's scale to make sure this type gets the same value for everyone, and then we shall have fully comparable scales.

If there is no type that is possible for everyone, the interpersonally comparable scale will have to be built by a sequence of pairwise comparisons. Each person can have her scale aligned with another person who can live a life of the same type as she can. We can hope that the whole population can be covered by overlapping pairs like this. If so, we can achieve fully comparable scales this way.

Since health is a component of temporal wellbeing, the interpersonal scale of health is a useful prototype. The scale of health used in qalys is built on a similar assumption. Let a 'health-type' of life be the set of features of a period of life that

[13] See Fleurbaey (2016), 'Equivalent income'.

contribute to determining how healthy a person is. We assume that two people are equally healthy if their lives share the same health-type. This makes levels of health interpersonally comparable.

Aligning people's scales in the way I have described is a way of adjusting each person's money values – her willingnesses to pay – according to the value of money for the person. All of a person's money-values are adjusted, with means that all her relative values remain the same. The value of extending her life relative to other goods such as food remain the same.

If we average in some way across the populations of different countries, the adjustments I have described give us an exchange rate between the countries' currencies. Let us call it the 'value parity' rate. The rupee/dollar value parity rate will be much higher than the rupee/dollar purchasing-power parity rate. Purchasing-power parity makes the rupee price of goods the same as their dollar price. But since people in India are poorer than people in the US, goods are more valuable to them. So purchasing-power parity undervalues Indian people's goods, including their lives. But at the value parity rate, equally good lives will be accorded equal value.

A very crude, simplified version of this proposal treats all types of life as having the same value. It assigns one particular value to every life year, the same for each person. Between countries, exchange rates will be set to make this so. The life year serves as a numeraire. This is plainly a very rough approximation, since not all life years actually have equal value. However, it probably approximates the truth much better than assuming all dollars have equal value – the assumption implicit in traditional cost-benefit analysis.

So even the crudest, simplest version of my suggestion will lead to better cost-benefit analysis than the traditional method.

References

Boadway, Robin, 'The welfare foundations of cost benefit analysis', *Economic Journal*, 84 (1974), pp. 926–39.

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

Diamond, P. A., 'Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility: comment', *Journal of Political Economy*, 75 (1967), pp. 765–6.

Drèze, Jacques, (1962) 'L'utilité sociale d'une vie humaine', *Revue Française de Recherche Opérationelle*, 23 (1962) pp. 93–118.

Fleishman, A. B., and M. J. Clark, 'Evaluating future detriment from radioactive discharges: judgements and implications for optimisation of protection', in *National Radiological Protection Board Report* NRPB–R132 (1982).

Fleurbaey, Marc, 'Equivalent income', in *The Oxford Handbook of Well-Being and Public Policy*, Oxford University Press, 2016, pp. 453–75.

Gorman, W. M., 'The intransitivity of certain criteria used in welfare economics', *Oxford Economic Papers*, 7 (1955), pp. 25–35.

Hicks, J. R., 'The foundations of welfare economics', *Economic Journal*, 49 (1939), pp. 697–712.

Kaldor, Nicholas, 'Welfare propositions of economics and interpersonal comparisons of utility', *Economic Journal*, 49 (1939), pp. 549–52.

MacKillop, Eleanor, and Sally Sheard, 'Quantifying life: Understanding the history of Quality Adjusted Life Years (QALYs)', *Social Science and Medicine*, 211 (2018), pp. 359–66.

Scitovsky, Tibor de, 'A note on welfare propositions in economics', *Review of Economic Studies*, 9 (1941), pp. 77–88.

Tim Campbell[1]

DALYs and the Minimally Good Life[2]

Climate change is expected to impact the health of current and future generations. For example, rising temperatures are expected to increase rainfall in some areas of the global south, thereby increasing the incidence of waterborne diseases, such as cholera. As the negative health impacts of climate change become more salient, so does the question of how to measure these impacts. The standard practice is to use Disability-Adjusted Life-Years (DALYs) as a proxy for health losses due to all disease causes, where one DALY represents the loss of the equivalent of one life year at full health. A natural thought is that governments and NGOs responsible for promoting and protecting people's health should use their resources in such a way as to maximize expected DALYs averted per dollar spent. But this may conflict with a *sufficientarian* view that many find attractive, namely that when it comes to global health, our priority should be to come as close as possible to a state in which every person can lead a life that is sufficiently good, a *minimally good life*. The potential conflict between maximizing expected DALYs averted and sufficientarianism sometimes goes unnoticed by ethicists focused on global health. This paper draws attention to the conflict by scrutinizing Nicole Hassoun's recent book *Global Health Impact*. It first explains the conflict, and then gives some reasons for being skeptical of sufficientarianism.

[1] Institute for Futures Studies, Stockholm, timothy.campbell@iffs.se

[2] Funding from Riksbankens Jubileumsfond (grant number M17-0372:1) is gratefully acknowledged. I am thankful to Nicole Hassoun and Anders Herlitz for helpful discussions of the topics in this paper.

1. Introduction

Climate change is expected to impact the health of current and future generations. For example, rising temperatures are expected to increase rainfall in some areas of the global south, thereby increasing the incidence of waterborne diseases, such as cholera. As the negative health impacts of climate change become more salient, so does the question of how to measure these impacts. The standard practice is to use Disability-Adjusted Life-Years (DALYs) as a proxy for losses due to all disease causes, where one DALY represents the loss of the equivalent of one life year at full health. A natural thought is that governments and NGOs responsible for promoting and protecting people's health should use their resources in such a way as to maximize expected DALYs averted per dollar spent. But this may conflict with a *sufficientarian* view that many find attractive, namely that when it comes to global health, our priority should be to come as close as possible to an outcome in which every person can lead a life that is sufficiently good, a *minimally good life*. The potential conflict between maximizing expected DALYs averted and sufficientarianism sometimes goes unnoticed by ethicists focused on global health. This paper first explains the conflict, and then gives some reasons for being skeptical of sufficientarianism. It argues that a criterion recently proposed by Nicole Hassoun for identifying the minimally good life fails because no life satisfies the conditions stated in the criterion.

In Section 2, I explain how ranking health interventions in terms of expected DALYs averted does not need to reflect how close those interventions bring us to the ideal that is central to sufficientarianism, namely an outcome in which every person can lead a minimally good life. In Section 3, I consider Hassoun's criterion for determining whether a life is minimally good, and I argue that no life could satisfy this criterion. While it is possible that someone will propose a better criterion, I suggest that the lack of any clear criterion casts doubt on sufficientarianism. We have reason to think that insofar as our maximizing aims are in tension with our sufficientarian aims, the latter should yield to the former. Section 4 summarizes the main points of the paper.

2. DALYs and Sufficientarianism

A ranking of health interventions in terms of DALYs averted will be useful only if that ranking approximates some other underlying ranking, for example the true *betterness* ranking, or *choiceworthiness* ranking, of health outcomes. A ranking of interventions in terms of DALYs averted should reflect our underlying axiology for health outcomes—our ordering of health outcomes in terms of betterness or choice-

worthiness. Our axiology need not be fully general. It could be tailored to a specific dimension that we consider important. If we are sufficientarians, then the relevant dimension might be people's ability to live a minimally good life. With a sufficientarian axiology in hand, we would be in a position to see how well a ranking of interventions in terms of the number of DALYs averted reflects the underlying 'true' ranking of outcomes with respect to people's ability to live a minimally good life. We may take the following as an example:

> *A sufficientarian axiology for health outcomes*: One health outcome *A* is better than another *B* with respect to people's ability to live a minimally good life *iff* A has a greater sum of weighted individual benefits than B, and A is equally as good as B *iff* A and B have equal sums of weighted individual benefits, where benefits are weighted such that they have (i) positive value when they fall below the threshold *L* for a minimally good life, (ii) positive value when they fall above *L*, and (iii) greater value the further below *L* they fall. Moreover, (iv) the positive value of any benefit that falls below *L* is greater than the positive value of any benefit that falls above *L*. (For comparison, see, e.g. Holtug, 2010: 226–227).

The idea here is that benefits to those would otherwise have lives below L are *superior* in terms of value to benefits to those who would otherwise have lives at or above L, and any benefit of the former kind matters more the further below L the beneficiary's life would otherwise be. Now whether, and to what extent, a ranking of health interventions in terms of DALYs averted provides a good approximation of the ranking of the resulting health outcomes according to a sufficientarian axiology depends crucially on what L is. Depending on what L is, there are different ways that the DALYs averted ranking might fail to reflect the sufficientarian ranking. In what follows, I shall first provide two rather abstract examples of the divergence of these rankings. Then, I will offer a more concrete example that might, depending on the specification of L, be an instance of one or the other of these abstract examples.

The first abstract example involves two health interventions, H1 and H2. H1 ranks higher than H2 in terms DALYs averted, but the benefits secured by H2 go to those who would be well below L in a no-treatment baseline, whereas the benefits secured by H1 go to those who would be above L in a no treatment baseline. Then, the outcome of H2 would have a greater sum of weighted benefits than the outcome of H1, and hence would be better than the outcome of H1 according to sufficientarianism. In this case, H1 is superior in terms of DALYs averted but H2 is superior in sufficientarian terms. The second abstract example involves two other health interventions, H3 and H4. H3 ranks higher than H4 in terms of DALYs averted. All the benefits secured by either of these interventions would go to those who would

be below L in a no treatment baseline. But the people benefited by H4 would be further below L in a no treatment baseline than would the people benefited by H3. For this reason, H4 produces a greater sum of weighted benefits than H3. In this case, H3 is superior in sufficientarian terms.

Without a specification of L, it will be difficult to point to more concrete instances of the two abstract examples just described. Here, I shall introduce an example that *might* count as an instance of one or the other of these abstract examples, but my aim in introducing the more concrete example is mainly illustrative.

Consider the Global Health Impact initiative's Disease Index, which assesses aggregate need for treatment for a range of different diseases.[3] In addition to diseases such as Malaria, HIV, and Tuberculosis, the index includes neglected tropical diseases such as Schistosomiasis. There is some evidence, although it is not definitive, that certain deworming interventions targeted at preventing neglected tropical diseases, including Schistosomiasis, are quite cost effective (Baird et al., 2016). It is not unrealistic that a certain hypothetical deworming intervention would avert more DALYs in expectation than some other hypothetical intervention that aims, for example, to prevent premature death from malaria. Call the first of these hypothetical interventions *the deworming intervention* and the second *the antimalaria intervention*.

However, it is also easy to imagine that the outcome of the deworming intervention would not be as good as that of the antimalaria intervention according to a sufficientarian axiology. One possible reason for this is that if the deworming intervention is not carried out, those who would otherwise have benefited from it will have lives that are at least minimally good, while, if the antimalaria intervention is not carried out, those who would otherwise have benefited from it will have lives that do not meet the standard of minimal goodness. This is at least somewhat plausible. Malaria kills many more people than Schistosomiasis, which disables more than it kills, and the majority of those who die from malaria die quite young (between the ages of 0 and 5) while those who die from Schistosomiasis typically die much older, although those with Schistosomiasis who do not die from it often suffer negative impacts on quality of life (Anisuzzaman and Tsuji, 2020; Simões et al., 2020; WHO, 2020). If there is a standard of minimal goodness for lives, it may be that most of those who die from malaria do not meet this standard—their lives may be far too brief. On the other hand, those who die from Schistosomiasis, or who suffer but do not die from it, may have lives that, while greatly negatively impacted, nevertheless meet the standard of minimal goodness. If this were true, then the deworming intervention and the antimalaria intervention may be like interventions

[3] Available at https://www.global-health-impact.org/index/disease/2015/summary#relocation_disease 2015).

H1 and H2 in my first abstract example; the former may avert more DALYs while the latter may be better according to a sufficientarian axiology because it produces a greater sum of weighted benefits.

But even if we assume that *all* Schistosomiasis victims in our example would have lives that fall short of the standard of minimal goodness (absent the deworming intervention), for the reasons just mentioned, it may be plausible to imagine that all the malaria victims in our example would have lives that fall *much shorter* of that standard (absent the antimalaria intervention). Hence, even if the deworming intervention and the antimalaria intervention are *not* like interventions H1 and H2 in my first abstract example, they could still be like interventions H3 and H4 in my second abstract example. In other words, it may be that all the benefits secured by either the deworming intervention or the antimalaria intervention would go to those who would be below the minimal goodness threshold L in a no treatment baseline, and yet, because the people benefited by the antimalaria intervention would be much further below L in a no treatment baseline than would the people benefited by the deworming intervention, the former would produce a greater sum of weighted benefits than would the latter.

The main takeaway of this section is that a sufficientarian axiology might favor an intervention that would, in expectation, avert fewer DALYs than another intervention. For this reason, there is an apparent philosophical tension between the sufficientarian aim of trying to ensure that everyone can have a minimally good life and the aim of producing the most, or the largest amount of, health benefit by maximizing DALYs averted. It is the first aim that we should pursue if, like Hassoun, we are sufficientarians. But, it is the second, maximizing aim that producers should have if they want to earn the right to carry the Global Health Impact label.

3. Specifying the Minimally Good Life

Is it likely that rankings of health interventions in terms of DALYs averted will disagree with the betterness ranking of outcomes that is given by sufficientarianism? We cannot give an answer to this question unless we can specify the minimally good life. Hassoun writes:

> On my account, to live a minimally good life, a person needs (1) *an adequate range* of the (2) *fundamental conditions* (3) *necessary and/or important for* (4) *securing* those (5) *relationships, pleasures, knowledge, appreciation, and worth- while activities, etc.* that (6) *a reasonable and caring person would set as a minimal standard of justifiable aspiration*. The relationships, pleasures, knowledge, appreciation, and worthwhile activities, etc. that a reasonable and caring person would set as a

> minimal standard of justifiable aspiration are *the things that make lives minimally good*. These things set the minimal standard to which people can justifiably aspire. (Hassoun, 2020: 19).

There are many questions one could raise about the different parts of this account. I focus on just one: How can we identify the minimal standard of justifiable aspiration that a reasonable and caring person would set? It may seem that if a life is at all worth living then one could justifiably aspire to have it. After all, a life worth living is a life that is *good* for the person who has it. But this is not enough according to Hassoun. She claims that a minimally good life must be better than a life that is only barely worth living. So the minimal standard cannot be identified as the standard of a life worth living.

Hassoun's answer is that the relevant standard can be identified using an empathy test (Hassoun, 2020: 22, n. 31). To tell whether a certain life is at least minimally good, ask yourself whether you would be *content* living that life. According to Hassoun, a person is content living a life when that person does not feel the need to change their situation when doing so is possible even at relatively low cost. Hassoun's idea, then, seems to be that if a reasonable and caring person would be content living a certain kind of life, then a life of that kind meets the minimal standard of justifiable aspiration; otherwise, it does not meet the standard.

The problem, though, is that whether I feel the need to change my situation when doing so is possible even at relatively low cost will depend crucially on what situations I imagine changing to when I consider the possibility of making such a change. A change is necessarily a change *from* one state *to* another. For any life situation *S* that I consider, I cannot say whether I would feel a need to *change* unless I am comparing *S* with at least one other life situation *S'*. For instance, consider a life of pure misery with no compensating good. If this were my life, I would not feel an unqualified need to change my situation. After all, some possible changes would result in a life that is even worse. I would feel a need to change my situation only if the change would give me something that is, to a sufficient degree, better than what I have.

But this is true for *any* possible life that I imagine. For instance, consider a fantastic life that involves living to age 100 completely free from disease and disability. We can also assume that this life contains an enormous amount of the goods that Hassoun lists in her point (5) above. If this were the default situation for me, would I feel the need to change my situation even at relatively low cost, assuming I could make such a change? Again, this depends on what I imagine that I would be changing *to*. Suppose it were possible for me to live an even better life. For example, suppose that there were some life-extending drug that would allow me to live 50 years longer

(to age 150) in perfect health. I may very well feel the need to acquire the drug, especially at relatively low cost, and so would not be "content" living without it. For any life that I consider, I find that I would feel a need to exchange it for a different life at relatively low cost, provided that the alternative life is, to a sufficient degree, better. All of this suggests that Hassoun's empathy test needs to be further refined if it is to pick out any life, or set of lives. But the refinement would need to specify not only the life that the caring and reasonable person must consider when deciding whether she feels the need to change, but also the alternative life, or lives, that she must consider changing *to*. Yet, there does not appear to be any non-arbitrary way of doing this.

A final complication that I will mention is that, even if we believe that there is such a thing as the minimally good life, it is unclear whether the standard for a minimally good life should be taken to apply to *whole lives* or to *parts* of lives. Some of Hassoun's remarks suggest the latter. For instance, she says 'the right' to a minimal decent life 'requires helping people live out their normal life expectancy at any given time minimally well' (Hassoun, 2020: 22). The qualification 'at any given time' suggests that we cannot just look at whole lives in order to determine whether a person meets the relevant standard. Even if a person has lived a fantastic life, if this person's final year of life was bad (because the person's health deteriorates in the final year), the person's right to a minimally good life would not be fulfilled (assuming her health fails because, for example, she was cut off from access to certain key medicines). This suggests that it may be quite difficult to satisfy a person's right to a minimally good life. We must ensure not only that people have the ability to live minimally good lives over- all but also that they are able to avoid dropping below the standard of a minimally good life at any given time. This also raises questions about, for example, what our sufficientarian axiology should say about parts of lives. Do we need an axiology that aggregates benefits across time-slices of lives rather than across whole lives, or perhaps one that aggregates across both time-slices and whole lives?

4. Conclusion

I have argued that there is a philosophical tension between Part 1 and Part 2 of Hassoun's project in *Global Health Impact*. The aim of sufficientarianism, applied to the sphere of global health, is not necessarily the aim of maximizing DALYs averted, and may even require abandoning this aim in some cases. I am not sure whether, or how much, the tension between Part 1 and Part 2 matters for practical purposes. This depends on where we set the threshold for a minimally good life. But the empathy test that Hassoun offers for identifying the minimally good life seems not

to identify any life. The test requires one to imagine the life situation of a person and determine whether one would feel the need to change from that situation to something else. But it matters what the 'something else' is. Whether one thinks that one would feel the need to change one's situation depends on what one imagines changing *to*. But for any life situation one imagines, one can imagine a better situation that one would feel the need to change to if one could. At least, this is true if one always prefers having more good life to having less good life.

References

Anisuzzaman and Tsuji, N. (2020). 'Schistosomiasis and Hookworm Infection in Humans: Disease Burden, Pathobiology and Anthelmintic Vaccines'. *Parasitology International*, **75**, 102051.

Baird, S., Hicks, J. M., Kremer, M., and Miguel, E. (2016). 'Worms at Work: Long-run Impacts of a Child Health Investment'. *The Quarterly Journal of Economics*, 131, 1637–1680.

Hassoun, N. (2020). *Global Health Impact: Extending Access to Essential Medicines*. New York: Oxford University Press.

Herlitz, A. (2019). 'The Indispensability of Sufficientarianism'. *Critical Review of International Social and Political Philosophy*, 22, 929–942.

Holtug, N. (2010). *Persons, Interests, and Justice*. New York: Oxford University Press.
Shields, L. (2016). *Just Enough: Sufficiency as a Demand of Justice*. Edinburgh: Edinburgh University Press.

Simões, T., Sena, R., and Meira, K. (2020). 'The Influence of the Age-period-cohort Effects on the Temporal Trend Mortality from Schistosomiasis in Brazil from 1980 to 2014'. *PLoS One*, **15**, e0231874. doi:10.1371/journal.pone.0231874.

World Health Organization. (2020). *World Malaria Report 2020: 20 Years of Global Progress & Challenges*. Geneva: World health Organization. Licence: CC BY-NC-SA 3.0 IGO.

Joe Roussos[1]

Uncertainty Attitudes as Values in Science[2]

There is now a large literature on values in science, discussing whether and how science can be objective while realistically acknowledging and managing the impact of values in the production of scientific information. In this paper, I am concerned with what counts as a value in this literature. Although previous discussions have identified a great many locations where value judgements occur, and a great many kinds of "inductive risk", they have nevertheless focused on a particular kind of evaluation. I call these evaluations of concrete outcomes. I argue that philosophers interested in values and science ought to additionally consider scientists' attitudes to uncertainty, which are evaluations of decision situations rather than concrete outcomes. I will be concerned with inductive risk, and the claim I make is a conditional one: if you are concerned about inductive risk in a particular part of science, then that concern should include uncertainty attitudes alongside the more commonly considered moral, social, or political values.

[1] Institute for Futures Studies, Stockholm, joe.roussos@iffs.se

[2] Funding from Riksbankens Jubileumsfond (grant number M17-0372:1) and the Global Challenges Foundation is gratefully acknowledged.

1. Introduction

There is now a large literature on values in science, discussing whether and how science can be objective while realistically acknowledging and managing the impact of values in the production of scientific information. In this paper, I am concerned with what counts as a value in this literature. I argue that we ought to consider scientists' attitudes to uncertainty as values. I will be concerned with inductive risk, and the claim I make is a conditional one: if you are concerned about inductive risk in a particular part of science, then that concern should include uncertainty attitudes alongside the more commonly considered moral, social, or political values.[3]

Here is a decision-focussed presentation of the argument from inductive risk (IR), and some major entries in the debate about it. The argument goes like this: Science involves decisions—about questions, methods, analyses, representations, uncertainty management, and whether to accept hypotheses, to name but a few. Decisions are a function of beliefs and desires, or to use language more familiar in this context, evidence and values. The value-free idealist hopes that these decisions can be made using only harmless epistemic and cognitive values. Rudner (1953) argues that the decision to accept or reject a hypothesis must involve moral evaluations of the badness of potential errors. Jeffrey (1956) argues that there is not one set of consequences for the scientist to consider but many, one for each application of the hypothesis. Thus, scientists should avoid acceptance decisions and merely report their probabilities for the hypotheses in question. Douglas (2009) responds that scientists face crucial decision points prior to the formation of the probabilities which guide the acceptance decision, so that Jeffrey's response is insufficient to remove values from science.

I focus on decisions because the core move of this essay is to note that decisions in fact depend on more than just beliefs and desires, they also depend on the decision maker's attitudes to uncertainty. The most familiar attitude to uncertainty is risk aversion, which I will discuss below alongside its cousin ambiguity aversion. I argue that these are evaluative attitudes, that they are not plausibly epistemic or cognitive values, and that they raise the same worries that motivate for the value-free ideal.

I will be concerned with two motivations for value-free science. The first reason is that moral values are thought to interfere with the pursuit of the core epistemic and cognitive goals of science: true theories, offering explanations and understanding of the world. If scientists are concerned with moral values, their scientific products might deviate from the truth. One way that this concern is presented is in

[3] For brevity, I will simply refer to "moral values", meaning that term to encompass the wide range of values discussed in this literature.

terms of "wishful thinking": a scientist whose concern for wellbeing informs their choice of methods, or acceptance of theories, might come to hold beliefs based on what they desire the world to be like rather than based on what it is really like. This argument has been associated with very strong forms of value-freedom, such as a complete absence of moral evaluations in the core practice of science.

The second reason is democratic. In order to be democratic, decisions must respond appropriately to the values of the people (represented by policymakers, political decision makers, etc.). Scientific input is required in policy, as we want our policies to respond to how the world really is. If science is value-laden then it "bakes in" some value judgements which are not those of the people, and these can influence which policy decisions are made. This subverts democratic control, as scientists' values dilute or supplant the values of the people. The result is what is sometimes called "liberal epistemic division of labour"—a picture of science-based policy according to which scientists provide the facts, and politicians provide the values (Brown 2009). Here, the relevant ideal is freedom from values which would interfere with democracy.

So, my main claim is that, insofar as moral values are a problem for these reasons, so are attitudes to uncertainty. Or, if these are not problems per se but rather something to be managed properly, then so too are uncertainty attitudes.

2. Values in the inductive risk literature

The values that are discussed in IR debates have a particular form. They take as their objects concrete outcomes—states of the world described without reference to probabilities. This is contrast with attitudes to uncertainty, which take as their object the state of uncertainty under which a decision is made.

Allow me to illustrate. In Rudner's (1953, 2) classic article, he argues that scientists "must make the decision that the evidence is sufficiently strong or that the probability is sufficiently high to warrant the acceptance of the hypothesis" under consideration. He presents examples which illustrate that this judgement of sufficiency is a moral matter. The first example contrasts a case in which "the hypothesis under consideration were to the effect that a toxic ingredient of a drug was not present in lethal quantity" with one where the "hypothesis stated that, on the basis of a sample, a certain lot of machine stamped belt buckles was not defective". He notes that the former would require higher confidence, in virtue of the high ethical stakes. We are, I take it, invited to imagine that the scientists must consider the harms to innocent medicine takers, weighing the badness of their illness or death against the potential benefits of treating the disease. This is in contrast to hapless customers whose trousers don't quite stay up.

Or, to take a more recent example, consider Winsberg, Oreskes and Lloyd's (2020) discussion of the science of extreme weather event attribution. Here, scientists attempt to determine whether a catastrophic event like Hurricane Harvey was due to or made worse by climate change. There is an ongoing methodological debate in this field between the so-called storyline approach and the more dominant fraction of attributable risk approach. Defenders of the storyline approach are "concerned that the risk-based approach will falsely fail to attribute the extreme event to climate change. [They are] concerned that this approach has a propensity to underestimate harm." Harm is here a straightforwardly moral matter: harm to people and society due to climate change and extreme weather events. The risk-based folks are "concerned about the risk of overstatement of human effects... [i.e.,] about making too many false positive errors, or overstating the role of climate change." These concerns are practical: "time and money might be spent preparing for events that will not occur" (Winsberg, Oreskes, and Lloyd 2020, 145–46).

The object of evaluation in each case is a state of the world which occurs after the decision and which are described without reference to the probabilities governing the decision. They are possibilities described in terms of the morally relevant facts, such as the harms of extreme weather events or the damage to the reputation of the scientists.

To make this more precise, let me present IR formally in terms of a stylised example of a decision to accept an hypothesis. (The presentation is Bayesian, so I use "accept"/ "reject" rather than the more familiar frequentist framing in terms of rejecting a null hypothesis.) Table 1 displays such an acceptance decision. Acceptance decisions are often framed in terms of the threshold of probability θ which is required to accept the hypothesis. In this framing, the object of interest is the value of θ, the idea being that only when $P(H) > \theta$ should the scientist accept H. The argument from IR says, roughly, that θ depends on (non-epistemic) evaluations of the consequences of that decision—represented by the utilities of outcomes in the decision table. E.g., $u(FP)$ represents the value of a false positive.

Table 1: An acceptance decision

	$P(H)$	$P(\neg H)$
Accept H	$u(TP)$	$u(FP)$
Reject H	$u(FN)$	$u(TN)$

I want to present the situation a little differently. Suppose that we have two scientists, who face a similar decision about accepting hypothesis H. They have the same

evidence, which they have evaluated identically, and thus they assign the same probability to H, say 0.6. They are faced with the decision to accept H, following Douglas (2009), let us suppose that they have first applied their epistemic and cognitive values. Nonetheless, a gap remains, which must be bridged by their evaluations of the consequences of error. Here, they differ; let us imagine that Scientist 1 takes a false positive to be neutral while Scientist 2 takes it to be moderately bad. The result is that the first scientist accepts hypothesis H and the second rejects it. Tables 2 and 3 display the decisions faced by each scientist and we can read off the third column that Scientist 1 accepts H while Scientist 2 rejects H.

The decision is guided by expected utility considerations: Scientist 1's expected utility of accepting is higher than that of rejecting. The point of the example decision tables is simply to make clear the nature of the evaluations and the role they are playing. The evaluation is of a consequence—a fully-specified state of the world which results if H is true (false) and if the scientist accepts (rejects). (This, not coincidentally, is how Jeffrey (1956) describes the consequences of choices.)

Table 2: Scientist 1's acceptance decision

	$P(H) = 0.6$	$P(\neg H) = 0.4$	
Accept H	2	0	$EU(A) = 1.2$
Reject H	1	1	$EU(R) = 1$

Table 3: Scientist 2's acceptance decision

	$P(H) = 0.6$	$P(\neg H) = 0.4$	
Accept H	2	-1	$EU(A) = 0.8$
Reject H	1	1	$EU(R) = 1$

3. Uncertainty attitudes and scientific decisions

The decision theoretic underpinnings of the above example are highly idealised. In particular, they ignore the fact that many actual agents have attitudes towards uncertainty itself. In simple terms, an uncertainty attitude is a liking of or aversion to uncertainty itself. An aversion to uncertainty manifests as a preference for making decisions with less uncertainty over making decisions with more uncertainty; more

subtly, such attitudes are measured via willingness to trade material consequences in exchange for a reduction of the uncertainty associated with making a decision. As I use the term here, "uncertainty attitude" is an umbrella category, which encompasses risk attitudes and ambiguity attitudes.

My argument for taking them seriously in the values and science debate is very simple: Many people have these uncertainty attitudes, and we should expect the same to be true of scientists.[4] These attitudes make a difference to the decisions people make—as I demonstrate below. They are plausibly rational.[5] So, a rational theory of scientific inference should account for them.

I begin with attitudes to risk. A decision-maker who is risk averse will prefer to receive 100 euros for sure than to place a wager which has 50% chance of paying 0 and 50% chance of paying 200. These bets have the same expected value in euros, so the decision maker's preference for the sure 100 is explained by their distaste for risk when it comes to getting euros. "Risk" here means the kind of uncertainty present in the wager: there are widely spread out outcomes which occur with known probabilities. The simple orthodox decision theory that I used above, expected utility theory, has room for only a limited kind of attitude to uncertainty. This is a form of risk aversion which can be captured in the shape of the agent's utility function: agents who are risk averse in euros are described with utility functions that are concave in euros. This means that they get more utility from the first 100 euros than they do from the second 100.

This way of representing risk averse behaviour can capture some real behaviour amongst people (and, I presume, scientists). But it is severely limited. The use of concave utility functions conflates two psychologically distinct phenomena: decreasing marginal utility of a good and an aversion to risk (Stefánsson and Bradley 2019). Expected utility theory also has no room for agents who are risk averse in utility itself—who would prefer a sure 100 units of the good over a 50-50 gamble of 0 and 200 units. Nevertheless, this is both possible and plausibly rational. In fact, there is significant evidence that agents have risk attitudes which cannot be represented via concave utility functions; indeed, such agents cannot be represented as maximising expected utility at all.[6] It is this kind of non-EU attitude that I refer to as "risk aversion" in the remainder of this essay.

For non-EU agents with uncertainty attitudes, these attitudes are an additional

[4] For empirical evidence and discussion of trends, see (Di Mauro and Maffioletti 2004) and (Trautmann and van de Kuilen 2015).

[5] As evidenced by the normative models of rational choice which account for them, e.g., (Buchak 2013) and (Stefánsson and Bradley 2019) for risk, and (Gilboa and Schmeidler 1989) and (Bradley 2017) for ambiguity.

[6] See the evidence referred to in footnote 4.

ingredient in their decision making, beyond their beliefs and desires. Agents who are risk averse are not merely responding to the expected value of their decisions, they are also responding to the fact that the outcomes are distributed a certain way, and that the component outcomes have the specific probabilities they do. Where the risk neutral expected utility maximiser regards all ways of getting an expected 100 utils as equivalent, the risk sensitive decision maker regards the sure 100 as different from the 50–50 bet on 0 and 200, and as different from the 1/3–2/3 bet on 0 and 150, and so on. Philosophers have defended this as a rational pattern of preference and offered models of such choices, e.g., the risk-weighted expected utility theory (REU) developed by Buchak (2013). Buchak's model contains a risk preference function r for each agent. Risk averse agents have convex risk functions, e.g., $r(p) = p^2$, while risk neutral agents have $r(p) = p$. This function modifies the probabilities, before they are combined with the agent's utilities in an expected value-type calculation. The details of the calculations don't matter; what does is that agents with identical probabilities and utilities can reach different decisions because they have different risk attitudes, represented by different r functions.

In canonical examples of IR, we suppose that scientists have some probabilities to hand and that they are deciding whether to accept a hypothesis on the basis of them—as in Table1. This is a straightforward example of a decision under risk. We then say that the scientist makes moral evaluations of the outcomes of error. Scientists, like other people, are plausibly risk averse when it comes to decisions about such values. So, imagine two scientists each confronted with the decision in Table 1, and suppose that they have the same evidence, same priors, and have arrived at the same probability for H. Suppose also that they make the same IR assessments: they identify all of the same consequences, and evaluate them identically. They agree, in other words, on exactly how bad each kind of error would be, and on how important true positives and true negatives are. It is nevertheless possible for these two scientists to make different inferences about H, because they differ in their taste for risk. So, since some scientific decisions involve risk, and since we should expect at least some scientists to have non-EU risk attitudes, we should expect these attitudes to be influencing the scientific information they produce.

The same thing can happen for attitudes to ambiguity. Ambiguity refers to a kind of uncertainty where we lack the information required to estimate probabilities precisely. Imagine a game based on drawing balls from an urn. Urn 1 has 50 red and 50 black. If it is black, there is no payment. Now consider Urn 2, which also has 100 balls in some unknown combination of red and black. A ball is drawn at random from the urn, and if it is red then the player is paid out 200. An agent is offered a choice: they can take draw from Urn 1 or from Urn 2. Knowing nothing about the distribution of balls in Urn 2, the decision-maker might as well take them to be equally likely,

but importantly they don't know that this is the case—unlike with Urn 1. A decision-maker who is *ambiguity averse* will prefer the bet on Urn 1, with known 50–50 odds, to a similar bet on Urn 2, with unknown odds.

Situations like this are sometimes represented with imprecise probabilities, e.g., a range of probabilities for H like 0.7–0.9. Decisions under ambiguity are controversial and, since most work on ambiguity has been done by economists and descriptive decision theorists, there is less in the way of normative theorising here. Nonetheless, philosophers like Bradley (2017) have defended the rational permissibility of attitudes to ambiguity and offered decision rules for different degrees of ambiguity aversion and ambiguity seeking. An example is the Alpha Maximin rule, according to which decision makers value options at a mixture of the worst and best expected utility, relative to the range of probabilities they entertain. (Alpha is the parameter controlling how much weight the worst-case EU gets.) Again, the precise details don't matter. What does is that with identical sets of probabilities and identical utilities can reach different decisions because they have different ambiguity attitudes, represented by different α parameters.

Many actual scientific decisions involve ambiguity; i.e., they are more like betting on Urn 2 than Urn 1. For example, recall Douglas' (2000, 571) case of scientists characterising unclear evidence. There, toxicologists examined rat liver slides under microscopes, looked for abnormalities, and classified them as benign or malignant. This is a skilful judgement, relying on experience and tacit knowledge. The classification of borderline cases is, Douglas argues, subject to IR. But it is not best characterised as a "decision under risk" in the technical sense described above since there no clear probabilities in play.

Scientists, like other people, plausibly have a range of attitudes to ambiguity. Consider a pair of scientists who face the same choice of whether to accept H. Let us suppose that there are different lines of evidence, perhaps gathered through different methods. Using each line of evidence, the scientists can assess the probability of H. Suppose again that their epistemic assessments of the evidence are identical. However, they don't know how to combine the different lines of evidence. So, they represent their uncertainty about H using the set of probabilities supported by the evidence. This is a form of imprecise probability, a common framework for representing ambiguity which is less demanding than the precise probabilistic framework used above. We again add the constraint that they make the same IR assessments: they identify all of the same consequences, and evaluate them identically. Nevertheless, one might accept H and the other reject it, because they differ in the attitudes to ambiguity. So, since some scientific decisions involve ambiguity, and since we should expect at least some scientists to have non-EU ambiguity attitudes, we should expect these attitudes to be influencing the scientific information they produce.

At this point, the reader might wonder whether this is really a new observation. Isn't the whole discussion about inductive *risks*, and attitudes thereto? Here it is worth noting that the English word "risk" is ambiguous between several meanings. It can mean an unwanted event, as in "the risk of getting cancer". It can mean the probabilities of unwanted events, as in "the risk that a smoker's life is shortened is 50%." In risk analysis, it almost always means the expectation values of unwanted events, as in "the risk of smoking is 12.5 life-years lost on average." Or, it can refer to spreads of outcomes over possibilities, as when I described the sure 100 euros as less risky than the 50–50 gamble on 0 or 200 euros. The IR discussion focusses on the value of the state of the world in which one has made the inductive error. This is a use of risk in the sense of bad outcome. But as we have seen, that is not what matters for non-EU uncertainty sensitive agents.

There are, to be sure, scattered references to risk aversion in the literature on values and science, for example the single unexplained mention by Reiss and Sprenger (2017). But it is not a central topic. It appears nowhere in Elliott and Richards (2017) anthology on inductive risk, not even in the wide-ranging typology of risk by J. B. Biddle et al. (2017). Nor does it appear in Biddle's earlier (2013) "state of the field" review. To some extent this is not surprising, as much of the action in the philosophy of risk and ambiguity attitudes is recent. When uncertainty attitudes appear, they do so in the context of recent discussions of science-driven policy decisions. For example, Bradley and Steele (2015) and Winsberg (2018) discuss the importance of the permissibility of uncertainty attitudes when discussing how policymakers can respond to the scientific uncertainty in climate science. They do not make the move I am making here, of recognising that those very scientists make IR decisions which themselves could permissibly have been affected by their attitudes to uncertainty.

4. Uncertainty attitudes matter

Now, drawing on Douglas (2000) and Winsberg (2018) I suggest that we see scientific information as the result of a series of decisions, possibly made by different scientists, with multiple points for IR-considerations to enter. We have just seen that uncertainty attitudes can alter how scientists make IR-sensitive decisions. Since any scientific product is the result of a sequence of such decisions, uncertainty attitudes could have a significant cumulative effect, leading scientists (or groups of scientists) who differ only in their attitudes to uncertainty to arrive at quite different conclusions at the end of the sequence.

So, they make a difference. Does this difference matter? One way we can think about this is to consider directly what kind of evaluation is happening when a scientist's attitude to uncertainty affects their decisions. On an illuminating account

due to Bradley and Stefánsson, uncertainty attitudes involve conative attitudes towards chances (Stefánsson and Bradley 2015, 2019; Bradley 2016). For example, risk seeking agents enjoy *facing risks*: for an experienced mountain climber "there is an optimal region of risk, where the chances of death or injury are high enough to require courage of the climber but not so high as to make the activity foolish" (Stefánsson and Bradley 2015, 605). It is not merely the concrete outcomes which are evaluated, but the chances of those outcomes. To take a more morally loaded example, one might think that scarce medical resources should be distributed by lottery because there is *value in having a chance* of receiving the treatment, even if one doesn't get it in the end. Clearly, evaluating chances-of-outcomes is related to an evaluation of the outcomes themselves—it is the harm of death that creates the thrill for the climber, and the benefit of life-saved that creates the value of the lottery. So, an uncertainty attitude in a decision situation depends on the underlying evaluation of the consequences, though it is not reducible to them (as the risk seeking mountain climber demonstrates). In other words, scientists' uncertainty attitudes are a distinct avenue through which their moral values affect their decisions.

Now, this is debated, so I will also offer two more direct considerations. First, I think there is a clear prima facie case that attitudes are a challenge to objectivity in the same way that moral values are. Consider the two cases side by side. Suppose that a scientist, Karim, wants to make use of hypothesis H in his work. To establish whether he should rely on H, he reads two scientific papers. Suppose that they emerged from the same lab and so the authors had the same prior empirical beliefs, collected exactly the same data, and evaluated it identically. Nevertheless, the authors reach different conclusions. Karim discovers that the difference is due to how they morally evaluated the badness of the potential inferential errors. To many, this dependency seems bad for Karim's project of learning facts about the world. Consider a variation of the story where the scientists also report, honestly let us suppose, identical moral values. They are upfront about their IR reasoning, and it is clear that they identified all the same outcomes and evaluated them identically. Nevertheless, the authors reach different conclusions. Karim investigates, and discovers that the difference is due to a difference in their risk attitudes: one is moderately risk averse, the other moderately risk seeking. If you are concerned by the first story, I find it hard to see how you can avoid being concerned by this variation.

A second concern is especially pressing for those whose preferred response to the value-ladenness of science is a form of democratic procedural solution in which stakeholder values are incorporated into science. The idea behind such approaches, as I understand them, is that by eliciting and incorporating stakeholder values scientists come to act on behalf of the non-expert users of science, so that the

science reflects what they would conclude if they were in the scientist's epistemic position. For example, discussing uncertainty management in climate adaptation work, Parker and Lusk (2019, 1647) write:

> [I]f choices must be made, they could be made in light of the [IR] preferences of the user or client: if it would be particularly bad for the user's purposes for uncertainty to be underestimated, then the provider might select those methodological options that will deliver a broader uncertainty estimate.

The key point once again is that decisions aren't only a function of evidence and values. These are filtered through attitudes to uncertainty when a decision is made. So if one attempts to incorporate stakeholder values into science while neglecting stakeholder uncertainty attitudes, this will lead to decisions which still diverge from how stakeholders would make them if they were in the scientist's epistemic position.

5. Potential objections and replies

There are several ways one could respond to my claim that we should worry about uncertainty attitudes whenever we worry about inductive risk.

A first objection might be that there is an obvious solution: each attitude-type discussed comes with a "neutral" variant, risk neutrality and ambiguity neutrality. Surely these are the attitudes required for value-neutral science. However, this would be to be misled by the names. There is nothing "neutral" about risk neutrality, in the sense typically meant be value-free idealists. Risk neutrality is just one of many evaluative stances; the one according to which (€100for sure) is equal in value to (€0, 0.5 ; €200r, 0.5). The strong value-free ideal is that scientists should make no moral evaluations whatsoever, and weaker forms distinguish particular kinds of problematic judgements, such as those which interfere with democracy. Clearly risk neutrality is not suitable for strong value-freedom. Now, one might argue that there is a risk attitude which does not interfere with democracy, and further argue that it is risk neutrality. But linguistic coincidence cannot supply that argument.

A second objection might be that uncertainty attitudes are harmless, as the values they encode are one of the kosher varieties, either cognitive or epistemic values. Let's begin with cognitive values. As I have used the term, these are properties of theories and models, which facilitate scientific cognition (following Douglas 2009, 93–94). That is to say that they facilitate thinking with and understanding these scientific theories and models, for agents with our cognitive capabilities. For example, simplicity "is a cognitive value because complex theories are more difficult

to work with, and the full implications of complex theories are harder to unpack" (Douglas 2009, 93). Uncertainty attitudes don't seem to fit the type: they aren't properties of theories or models, and achieving the preferred valence doesn't facilitate cognition. The uncertainty averse agent favours decision situations in which the harms and benefits are "clumped up" rather than "spread out" across the possibilities that they are aware of. But the benefit isn't cognitive. The risk averse agent has no trouble reasoning about spread out outcomes—indeed, they do so as part of their decision-making.

Perhaps then uncertainty attitudes are of the pure epistemic type. These, recall, constitute the truth-seeking mission of science. Whereas cognitive values can have little to do with truth seeking or truth preservation (Laudan 2004), epistemic values are tightly focused on these aims. On this categorisation, the epistemic is a small category containing values such as predictive accuracy, internal consistency, and empirical adequacy. The initial plausibility of including uncertainty attitudes is therefore quite low. Uncertainty attitudes are towards decision situations and patterns of outcomes of choices, not to states of knowledge. We can see this most clearly in the famous proofs that risk averse agents, who appear to value certainty, can sometimes rationally turn down free information (Wakker 1988; Buchak 2010; Campbell-Moore and Salow 2020).

But we can do better than a mere argument by elimination of alternatives, even if we don't accept Bradley and Stefánsson's account of the evaluative basis of uncertainty attitudes. There are independent reasons to think that uncertainty attitudes are a moral matter, which come to us from the ethics of social decision-making. There is now a small literature on whether the uncertainty attitudes of social decision makers matter to the ethics of their decisions. For example, in a recent manuscript Buchak (ms) argues that specific attitudes are morally required: social decision makers should be risk averse but ambiguity neutral, unless they know the attitudes of every person on whose behalf they decide, in which case they should defer to those. Stefánsson (forthcoming), without relying on his values-of-chances account, argues that social decision makers ought to be more risk seeking than individuals would be if they were making the decisions themselves. Rowe and Voorhoeve (2018) argue that ambiguity aversion is permissible for social decision makers, and that this fact supports a form of egalitarianism. Now recall that, on the democratic motivation for the value-free ideal, an important use of science is that it informs policy. Here, scientists are involved in decisions made on behalf of others, and so the above gives us an angle on the moral status of their attitudes. The conclusion I want to draw from this is that, at minimum, uncertainty attitudes matter to moral decision making. So, the uncertainty attitudes of scientists may interfere with democratic decision making, just as their moral attitudes might.

6. Conclusion

Where does this leave us? It might be that there is a morally correct set of uncertainty attitudes for scientists to take. (They might even be the neutral attitudes.) Or there might be a set of attitudes which do not interfere with democracy. Or we could follow the turn in the values and science debate towards stakeholder engagement, and insist that science ought to be based on the "right values" in a procedural sense: values supplied to them as the output of a consultative procedure with some relevant group of users of the scientific outputs.

All of these are live options in the values and science debate. Moreover, they occur because of the ethical implications of uncertainty attitudes. This establishes my conditional claim: if you are concerned about inductive risk in a particular part of science, then that concern should include uncertainty attitudes alongside the more commonly considered moral, social, or political values.

References

Biddle, Justin. 2013. "State of the Field: Transient Underdetermination and Values in Science." *Studies in History and Philosophy of Science Part A* 44 (1): 124–33. https://doi.org/10.1016/j.shpsa.2012.09.003.

Biddle, Justin B., Quill Kukla, Kevin C. Elliott, and Ted Richards. 2017. "The Geography of Epistemic Risk." In *Exploring Inductive Risk: Case Studies of Values in Science*. Oxford: Oxford University Press.

Bradley, Richard. 2016. "Ellsberg's Paradox and the Value of Chances." *Economics and Philosophy* 32 (02): 231–48. https://doi.org/10.1017/S0266267115000358.

———. 2017. *Decision Theory with a Human Face*. Cambridge University Press.

Bradley, Richard, and Katie Steele. 2015. "Making Climate Decisions." *Philosophy Compass* 10 (11): 799–810. https://doi.org/10.1111/phc3.12259.

Brown, Mark B. 2009. *Science in Democracy: Expertise, Institutions, and Representation*. Cambridge, MA: MIT Press.

Buchak, Lara. ms. "Risk and Ambiguity in Ethical Decision Making." manuscript.

———. 2010. "Instrumental Rationality, Epistemic Rationality, and Evidence Gathering." *Philosophical Perspectives* 24 (1): 85–120.

———. 2013. Risk and Rationality. Oxford University Press.

Campbell-Moore, Catrin, and Bernhard Salow. 2020. "Avoiding Risk and Avoiding Evidence." *Australasian Journal of Philosophy* 98 (3): 495–515. https://doi.org/10.1080/00048402.2019.1697305.

Di Mauro, Carmela, and Anna Maffioletti. 2004. "Attitudes to Risk and Attitudes to Uncertainty: Experimental Evidence." *Applied Economics* 36 (4): 357–72. https://doi.org/10.1080/00036840410001674286.

Douglas, Heather. 2000. "Inductive Risk and Values in Science." *Philosophy of Science* 67 (4): 559–79. https://doi.org/10.1086/392855.

———. 2009. *Science, Policy and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Elliott, K, and T Richards. 2017. *Exploring Inductive Risk: Case Studies of Values in Science*. Oxford: Oxford University Press.

Gilboa, Itzhak, and David Schmeidler. 1989. "Maxmin Expected Utility with Non-Unique Prior." *Journal of Mathematical Economics* 18 (2): 141–53. https://doi.org/10.1016/0304-4068(89)90018-9.

Jeffrey, Richard C. 1956. "Valuation and Acceptance of Scientific Hypotheses." *Philosophy of Science* 23 (3): 237–46.

Laudan, Larry. 2004. "The Epistemic, the Cognitive, and the Social." In *Science, Values, and Objectivity*, edited by Peter Machamer and Gereon Wolters, 14–23. Pittsburgh: University of Pittsburgh Press.

Parker, Wendy S, and Greg Lusk. 2019. "Incorporating User Values into Climate Services." *American Meteorological Society* 100 (9): 1643–51.

Reiss, Julian, and Jan Sprenger. 2017. "Scientific Objectivity." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2017/entries/scientific-objectivity/.

Rowe, Thomas, and Alex Voorhoeve. 2018. "Egalitarianism Under Severe Uncertainty." *Philosophy and Public Affairs* 46 (3): 239–68.

Rudner, Richard. 1953. "The Scientist Qua Scientist Makes Value Judgments." *Philosophy of Science* 20 (1): 1–6.

Stefánsson, H. Orri. forthcoming. "Longtermism and Social Risk Taking." In *Essays on Longtermism*, edited by Barrett, Hilary Greaves, and David Thorstad. Oxford: Oxford University Press.

Stefánsson, H. Orri, and Richard Bradley. 2015. "How Valuable Are Chances?" *Philosophy of Science* 82: 602–25.

———. 2019. "What Is Risk Aversion?" *The British Journal for the Philosophy of Science* 70: 77–102.

Trautmann, S. T., and G. van de Kuilen. 2015. "Ambiguity Attitudes." In *The Wiley Blackwell Handbook of Judgment and Decision Making*, edited by Gideon Keren and George Wu, 89–116. John Wiley & Sons.

Wakker, Peter. 1988. "Nonexpected Utility as Aversion of Information." *Journal of Behavioural Decision Making* 1: 169–75.

Winsberg, Eric. 2018. *Philosophy and Climate Science*. Cambridge: Cambridge University Press.

Winsberg, Eric, Naomi Oreskes, and Elisabeth A. Lloyd. 2020. "Severe Weather Event Attribution: Why Values Won't Go Away." *Studies in History and Philosophy of Science Part* A 84: 142–49.

Karim Jebari[1] & Martin Kolk[2]

Sex Selection for Daughters: Demographic Consequences of Female-biased Sex Ratios[3]

Modern fertility techniques like flow cytometry allow parents to carry out preimplantation sex selection at low cost, no medical risk, and without the ethical or medical concerns associated with late abortions. Sex selection for non-medical purposes is legal in many high-income countries, and social norms toward assisted reproductive technology are increasingly permissive and may plausibly become increasingly prevalent in the near future. While concerns over son preference have been widely discussed, sex selection that favors female children is a more likely outcome in high-income countries. If sex selection is adopted, it may bias the sex ratio in a given population. Male-biased populations are likely to experience slower population growth, which limits the long-term viability of corresponding cultural norms. Conversely, female-biased populations are likely to experience faster population growth. Cultural norms that promote female-biased sex ratios are as a consequence therefore also self-reinforcing. In this study, we explore the demographic consequences of a female-biased sex ratio for population growth and population age structure. We also discuss the technology and parental preferences that may give rise to such a scenario.

[1] Institute for Futures Studies, karim.jebari@iffs.se

[2] Institute for Futures Studies, and Stockholm university, martin.kolk@iffs.se

[3] Funding from Riksbankens Jubileumsfond (grant number M17-0372:1) is gratefully acknowledged.

1. Introduction

Preimplantation sex selection (henceforth sex selection) is a practice in which individuals attempt to control the sex of their offspring before the fertilized egg has been implanted in the uterus. The motivation for sex selection can vary, but we will here focus on sex selection for non-medical reasons. In recent decades, post-implantation sex selection (i.e., abortions and infanticide) has contributed to bias in the sex ratio in certain countries; this phenomenon has been widely discussed (Sen, 1990). Most research on biased sex ratios at birth due to parental preferences has focused on "missing women," often with a focus on Asian countries where sex selection against female children and excess child mortality among girls have been prevalent (e.g. Guilmoto, 2012; Sen, 1990).

In this article, we discuss the potential impact of an increase in the use of sex selection technology from a different perspective, based on three recent developments:

- Sex selection technologies are now legal, non-invasive, and relatively inexpensive in many high-income countries.
- In high-income countries, prospective parents have on average, a preference for female children.
- In most high-income countries, single women[4] and women in same-sex relationships have unprecedented legal and financial access to assisted reproductive technology (ART).

We note that unlike a male-biased sex ratio, which tends to depress population growth (Johnson, 1994), a female-biased sex ratio will increase population growth (ceteris paribus). In this article, we explore the demographic consequences of female-biased sex ratios at birth and show that they may be considerable under certain assumptions.[5] We show how such a will affect number of births and population size (r in a demographic or population genetics model), through a cultural evolutionary process.

[4] By "single women," we mean women that choose to procreate and rear a child as the sole caretaking parent.

[5] Here, we quote the prescient foreword by Nathan Keyfitz to the pioneering edited volume (Bennet 1983) exploring potential future consequences of how sex selective abortion may give male biased sex ratios: "Too often we have to wait until an invention has been in use for a long time for social science to investigate and explain its effect. We are fortunate in this instance that a group has taken the initiative to start the social investigation before the invention comes to technical maturity and long before it is actually adopted"

2. Reproductive Technologies

Three techniques are currently used for the purpose of sex selection. The first two are relatively invasive, expensive, and associated with non-negligible medical risk. We mention these in contrast to the third technique.

Ultrasound in combination with abortion is a prenatal rather than a preimplantation technique for sex selection and is thus more invasive and associated with considerable medical risk. The sex of the fetus can be detected with ultrasound at week 11, at the earliest, which means that abortions may have medical risks, especially in low-income countries (Igbinedion & Akhigbe, 2012).[6] While this remains the most prevalent technique for sex selection in low- and middle-income countries (mostly used to select male children), it is rarely used in high-income countries. We do not foresee this as a common or preferred method for sex selection in high-income countries in the future; thus, we will not discuss this technique in any further detail.

Preimplantation genetic diagnosis (PGD), a practice where an embryo is screened in vitro before implantation in the uterus, is used to some extent for sex selection in high-income countries. PGD is highly accurate in determining the sex of the embryo (Harper & SenGupta, 2012; Sermon et al., 2004). However, since PGD requires in vitro fertilization (IVF), it is relatively expensive and invasive, as it requires hormonal ovarian stimulation and retrieval from the ovaries. Thus, PGD is typically motivated by medical sex selection: for example, if the parents have a hereditary medical condition that only affects one sex. However, as a larger share of parents use IVF for reasons other than sex selection (Kupka et al., 2014), more parents will be able, at little additional cost, to choose the sex of their child if they so desire.

Flow cytometry is a relatively novel technique that is far less invasive and costly than the other two. Here, semen is labeled with a fluorescent dye that binds to the DNA of each spermatozoon (Sharpe & Evans, 2009). As the X chromosome is larger (i.e., contains more DNA) than the Y chromosome, "female" (X-chromosome bearing) spermatozoa will absorb a greater amount of dye than their "male" (Y-chromosome bearing) counterpart. Consequently, when exposed to UV light, "female" spermatozoa fluoresce brighter than "male" spermatozoa. As the spermatozoa pass through the flow cytometer in single file, each spermatozoon is encased by a single droplet of fluid and assigned an electric charge corresponding to its chromosome status (X-positive charge or Y-negative charge). The stream of X- and Y- droplets is then separated using electrostatic deflection and collected into

[6] Although possible, such early attempts at sex determination are prone to a high degree of false negatives.

separate collection tubes for subsequent processing (O'Neill, 2013; Reubinoff & Schenker, 1996). This method does not require IVF and can be used in combination with insemination. While this method is less invasive than PGD, it is also (somewhat) less accurate. In a study from 2014, 95% of babies born were females after sorting for X- spermatozoa and 85% were males after sorting for Y-bearing spermatozoa (Karabinus et al., 2014). The technique is more accurate when selecting female children than male children (Karabinus et al., 2014).

Sex selection is legal and in use in some high-income countries, including the U.S. (Bhatia, 2018). However, the most prevalent technique for sex selection in high-income countries, PGD, requires IVF, and the extent to which this is available for non-medical purposes varies. The U.S. has a very permissive regulatory regime, allowing so-called "fertility tourism" from other countries where sex selection is only allowed for medical purposes (Whittaker, 2011). However, with the increasing popularity of flow cytometry, access to sex selection is also likely to increase. No high-income country has banned flow cytometry, and a ban on insemination of sorted sperm is likely to be difficult to enforce. The company MicroSort, which uses flow cytometry, and offers non-medical sex selection services, already operates in Mexico, Malaysia, North Cyprus, and Switzerland, attracting fertility tourism (MicroSort, 2020).

Since the first IVF procedure in 1978, ART has become widespread and widely accepted. In the U.S., more than 55,000 women per year give birth to a baby conceived through ART (IVF or insemination; Dusenbery, 2020). Moreover, public support for this technology has also increased considerably, with ART now subsidized by public healthcare systems in many high-income countries for infertile different-sex couples, single mothers, and female same-sex couples.

3. Sex Preferences in High-income and Middle-income Countries

Recent research in sociology and demography has found increasing preferences for female children in high-income countries. This has mostly been expressed through parents more often having higher-order births if their previous children were either lacking sons or daughters, but there is also increasing evidence for parents explicitly wanting daughters when they have more direct choice over their reproduction. Below, we summarize the recent research on sex preferences in high-income countries.

Most of the existing research on sex preferences and fertility outcomes has focused on countries with strong son preferences. In particular, in East and South Asia, where patrilineal kinship systems are common, parental preferences for male children have been commonplace in both contemporary and historical societies

(Arnold & Zhaoxiang, 1992; Drixler, 2013; Guilmoto, 2012; Mungello, 2008; Sen, 1990). Such preferences have historically been associated with elevated female child mortality and infanticide, with major demographic impact (Arnold & Zhaoxiang, 1992; Drixler, 2013; Guilmoto, 2012; Mungello, 2008; Sen, 1990). In the 1990s and 2000s, the availability of ultrasound combined with abortion led to elevated male sex ratios across East Asia, South Asia, and Caucasia (Guilmoto, 2009). A preference for male children has also been historically common throughout Western Europe, but with only limited effects on child mortality or fertility outcomes (Kolk, 2011; Sandström & Vikström, 2015; Tsuya et al., 2010).

While a preference for sons seems to be the major determinant of childbearing decisions globally (Arnold, 1997; Guilmoto & Tove, 2015), in high-income countries, the picture is notably different, with increasing evidence of a preference for daughters and for a mixed-sex composition (Kolk & Schnettler, 2013; Miranda et al., 2018). This trend is not only prevalent in Western countries but has also been observed in Japan (Fuse, 2013). Also, in middle-income countries traditionally dominated by strong son preferences, such as rural China, there is some novel evidence that some parents are developing a preference for daughters over sons (Shi, 2017). In high-income countries, sex preferences are not expressed through biased sex ratios at birth; however, the sex composition of previous children has a strong impact on parity progressions (the decision to have a subsequent child). Across Europe and the U.S., research has shown that transition to higher order births is influenced by the sex composition of previous children (Blau et al., 2019; Hank, 2007; Hank & Kohler, 2000). A pattern in which parents prefer children of each sex is increasingly common; it is strongest for the transition to a third child, where a transition to a third child is least prevalent for parents with a son and a daughter (Hank & Kohler, 2000; Miranda et al., 2018).

Interestingly, in Nordic countries, while a preference for mixed-sex composition remains the dominant pattern, evidence points to more parents displaying daughter preference over son preference (Kolk & Schnettler, 2013; Miranda et al., 2018). For parents with one child, 35% of parents who had a son preferred their second child to be a girl, whereas only 23.4% of parents who had a daughter preferred their next child to be a boy (Miranda et al., 2018). For parents with a daughter, 74% said the sex of the next child did not matter, compared to 58% for those with a son. Demographers have previously speculated that high gender equality would lead to parental sex indifference, whereby the sex composition of previous children would not affect the decision to have subsequent children (Pollard & Morgan, 2002). In reality, however, this seems not to be the case; instead, we find that in countries that are the most equal, it appears more common for parents to prefer female children (Andersson et al., 2006; Miranda et al., 2018). It is thus conceivable that increasing

gender equality will, if anything, lead to daughter preference becoming more widespread. Moreover, in Japan, traditionalism and adherence to traditional gender roles among women have also been predictive of daughter preferences (Fuse, 2013), suggesting that—with increasing female agency over fertility—we may also see greater daughter preference in less gender egalitarian contexts (see also Shi, 2017). It should be noted here that, while they are a clear marker of parental sex preferences and the distribution of sons and daughters within families, sex-biased parity progressions (as described above) do not affect the overall sex ratio in a population.

Evidently, although most heterosexual parents in Western countries have not acted to deliberately affect the sex of their children, their behavior following the (random) allocation of previous births has a clear impact on their subsequent behavior. In the less prevalent contexts where potential parents already have direct agency over the sex of their children, we find stronger evidence of parental preferences for daughters. For several decades, adoptive parents have, on average, shown a strong preference for female children, an interesting illustration of a scenario in which parents to some extent can choose the sex of their children (Högbacka, 2008). However, sex and other aspects, such as the ethnic match of the child and the parents, interact in complex ways in international adoption (Högbacka 2008). In a sample of infertile women considering ART treatment, 40% responded positively about choosing the sex of their child if the option to do so was offered at no additional cost (Jain et al., 2005). Among women who wanted to select the sex of their future child, 39% wanted a male child, and 61% wanted a female child (Jain et al., 2005). Of women considering ART treatment, it was much more common to express a daughter preference than a son preference, although women who already had children had a preference for a mixed-sex composition (Jain et al., 2005). Lamberts et al. (2017) found higher rates of vasectomy among men with more sons than daughters. Overall, it seems that when more choice, technology, and agency are associated with the process of having a child (as opposed to children conceived through intercourse in heterosexual unions), the more parents accept and consider the option of choosing the sex of their children. When parents explicitly consider the choice of sex of their future children, a daughter preference seems more common.

Most children are reared by different-sex couples. In the research on sex preferences of partnered men and women, prospective mothers are seen to have a relatively higher preference for female children, whereas prospective fathers have a relatively higher preference for male children (Higginson & Aarssen, 2011; Lynch et al., 2018). This sex-biased pattern is found in societies with both son preference and daughter preference on average (van Balen, 2006). This suggests that the preference for children of one's own gender is a relatively general pattern across cultures. There is little research knowledge about sex preferences among single women and same-

sex female couples, although some studies indicate that heterosexual single women and women in same-sex couples more often exhibit daughter preference (Gartrell et al., 1996; Goldberg, 2009; Leiblum et al., 1995). If women across all union types have a preference for daughters one would expect single women and female same-sex couples to be able to act on this preference without negotiating with a male partner and therefore on average engage more readily in sex selection. In general, groups who are more likely to use ART for non sex-selective reasons, such as people with fertility concerns, single women, and women in same-sex relationships, may more often choose sex selection, because sex selection (either via PGD or flow cytometry) is a relatively straightforward addition to ART procedures (van Balen 2006).

In summary, previous research has found increasing evidence of a daughter preference in high-income countries. In situations where parents have more direct choice over the sex of their children, such as adoption and IVF, we also find stronger daughter preferences. Overall, we argue that a latent daughter preference is apparent in high-income countries, and that this is stronger among women than men.

4. Results

To estimate how female-biased sex ratios may affect population growth, we present calculations for different countries with different sex ratios and show how the sex ratio affects population growth rates and age structure.

Our calculations are based on two important assumptions. First, since we are interested in the long-term effects of changing sex ratios, we therefore show the long-term equilibrium effect of a change in sex ratios given a set of assumptions on fertility and mortality. This is what demographers refer to as a stable population (Wachter 2014, 218-249). It is worth stressing that we apply reductionist and commonly used demographic approximations to our demographic examples so they may be more easily followed, instead of more technical models. All our calculations refer to the long-term consequences for a population with the same fixed behavior over multiple generations; as such, they are only useful to illustrate the long-term implications of female-biased sex ratios. They are not useful for predicting actual demographic outcomes in the near future. Given the uncertainties in the extent of uptake (and timing of uptake) of the behaviors we discuss here, focusing on the large-scale demographic influence of these trends is more relevant than trying to forecast near-future empirical scenarios.

Second, our models follow the standard demographic methodology in which demographic analysis is based on female reproductive choice in a society. This is the approach used in most standard demographic analysis (e.g., Wachter 2014, 79-89). However, certain assumptions in such models, such as the implicit assumption that

(male) co-procreating and co-childrearing partners are unconstrained, are less realistic in cases with highly biased sex ratios. We discuss whether these assumptions can be analyzed independently of the overall sex ratio in section 5.3, as well as other factors that may stabilize the sex ratio given a preference for female births. We also analytically calculate the age structure implications (section 4.3) of different fertility scenarios with different sex ratios, showing that in some scenarios of high total fertility and highly biased sex ratios, it is relevant to assess demographic support ratios. Some previous demographic literature on male-biased sex ratios has created demographic models exploring how male preferences and male sex ratios affect population growth (Leung, 1994; Bennet, 1983; Mason & Bennett, 1977), finding that male-biased sex selection would decrease population growth. All data and calculations are available in a spreadsheet (supplemental file 1: data and calculations).

4.1 Consequences of Biased Sex Ratios for Population Growth

We begin by examining changes to population growth arising from different assumptions of fertility and share of female births. In Equation 1, we show the net reproductive rate based on a given sex ratio ($B_f/(B_t)$, female births over all births), mortality pattern (l_x), and fertility pattern (f_x) for five-year life tables. The net reproductive rate can be interpreted as a multiplier of each subsequent generation; a value above 1 thus indicates a population where every generation is larger than the previous one, and a value smaller than 1 indicates a shrinking population. A value of 1.5 indicates that each new generation is 50% larger than the preceding one.

Equation 1:

$$\text{NRR} = 5\frac{B_f}{B_t}\sum_{x=15-19,by\ steps\ of\ 5}^{45-49} f_x\ l_x$$

In Equation 2, we show the Dublin-Lotka approximation, which shows how a given net reproductive rate translates into yearly population growth (r) based on the mean age of reproduction in a population (T), which can be calculated from f_x.

Equation 2:

$$\text{r} \simeq \frac{ln(NRR)}{T}$$

Together, these two equations give a good approximation of how a given sex ratio, mortality pattern, and fertility pattern jointly determine long-term growth in a

given population. We largely focus on sex ratios and fertility here, as different mortality assumptions make little practical difference to contemporary populations in high-income countries. We use a single mortality pattern in all examples, based on the pattern in the U.S. for 2017. Survival up to age 45 is today so high, even in lower-middle income countries,[7] that it plays only a minor role in generational reproduction, and we do not expect this to change in the foreseeable future.

We illustrate the consequences of changing sex ratios with a selection of different assumptions on the average number of births per woman in a population (or Total Fertility Rate, TFR). The different fertility scenarios are: very low (Taiwan 2014, TFR = 1.16); somewhat typical for an OECD country (U.S. 2017, TFR = 1.76); high (Kenya 2014, TFR = 3.90); and the traditional fertility schedule used for populations that are close to the highest observed fertility in human populations, the Hutterites in the U.S. in the 1920s (TFR = 10.31; see Henry, 1961, from where we get our fertility schedule). We collected mortality data for the U.S. (Human Mortality Database—U.S., n.d.), fertility data for U.S. and Taiwan (Human Fertility Database—U.S. & Taiwan, n.d.), and fertility data for Kenya based on the 2014 demographic and health survey (Human Fertility Collection—Kenya, n.d.).

In Table 1 below (upper panel), we show the consequences of female-biased sex ratios for population growth (r) using the approximations from equations 1 and 2 with different fertility rates and sex ratios. We show fertility patterns for three different countries. We show a sex ratio for 48 daughters from 100 total births (which is close to what is naturally occurring in contemporary populations; see James, 1987), as well as sex ratios of 60[8] and 80 daughters per 100 births.

Equation 3 shows how yearly population growth (r) corresponds to initial and final population size (P_0 and P_n, respectively) over n years. We use Equation 3 to translate how the population growth rate in Table 1 (lower panel) translates into population growth over 50 years. This represents how much larger a population would be after 50 years of corresponding population growth, given that the fertility rates, mortality rates, population growth, and sex ratios would be fixed and at equilibrium.

[7] Survival to age 45 for a woman is around 95% for the US life table in our calculations, and above 90% for a country like modern-day Indonesia.

[8] For example, one conceptual scenario of a sex ratio of 60 could arise from a stratified population where (a) out of parents with two children, 50% of those with two sons choose sex selection for family balancing, but only 25% of those with 2 daughters do the same; b) within a group of single women and female same-sex couples, 45% choose to have only daughters; and c) out of everyone else, 20% choose to have only daughters. The effect on the population sex ratio can be meaningful even if only a minority of the population choose to use sex selection technology.

Table 1. Consequences of Varying Assumptions of Fertility Rates, Sex Ratios, and Mortality for Population Growth Rate (upper panel) and Population Change over 50 years (lower panel).

Population growth rate (r)

Fertility schedule	Taiwan (2014)	U.S. (2017)	Kenya (2014)
Number of female births over 100 total births			
48	-0.0216	-0.0066	0.0217
60	-0.0136	0.0014	0.0296
80	-0.0034	0.0116	0.0398

Effect over 50 years (ratio to original population)

Fertility schedule	Taiwan (2014)	U.S. (2017)	Kenya (2014)
Number of female births over 100 total births			
48	0.336	0.720	2.921
60	0.503	1.071	4.300
80	0.843	1.781	7.049
Total fertility rate	1.16	1.76	3.90
Mean age of childbearing	31.08	29.39	28.13

Note. The table refers to the stable population equilibrium resulting from the different combinations of rates.

Equation 3:

$$P_n = (1 + r)^n \, P_0$$

As we can see from the table, changing the sex ratio has a significant impact on population growth.[9] While the prevailing sex ratio and fertility schedule in the U.S. will lead to a reduction in the population by around 30% (the equilibrium consequences of contemporary demographic rates over 50 years), when we compare this to a population where an assumed 60% of all births are female, the population would instead surge by 7%. In turn, a population where 80% of births are female would

[9] These numbers do not account for any demographic change related to migration, for example, or any forecasted change in demographic rates.

increase by 78% over 50 years. With prevailing fertility rates and an unbiased sex ratio, Taiwan's current fertility schedule implies that such a population will contract by 66% over 50 years; given a sex ratio of 80% women, it would only contract by 16%. In Kenya, a sex ratio of 80% (rather than 48%) would result in an increase of over 600% (compared to the already considerable increase of nearly 200%) over 50 years. Note once again that these are equilibrium scenarios used to explore differences for changing sex ratios and not as demographic forecasts; here, we use constant demographic rates and contemporary constant U.S. mortality patterns to illustrate how changing sex ratios interact with different fertility assumptions. In the online appendix S1, we explore other demographic scenarios including very skewed sex ratios, male-biased sex ratios and very high fertility rates.

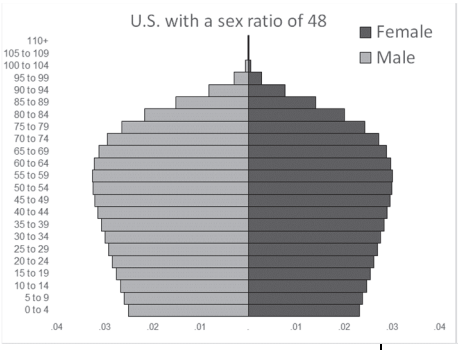
4.2 Consequences of Biased Sex Ratios for Population Age Structure

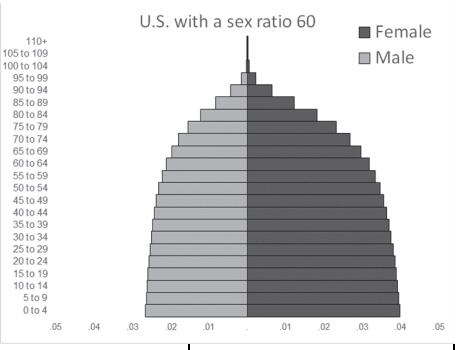
In the following section, we explore the consequences for the age structure of the population given the scenarios outlined above. In human societies, children and infants are provided resources and care by adults (to a large extent their parents), and in contemporary high-income societies, elderly individuals also provide substantial support. The importance of lifecycle transfers to the age structure of a population has been widely recognized in anthropology, economics, and evolutionary biology, where it forms the basis of life course theory (Kaplan, 1994; Lancaster et al., 2000; Lee & Mason, 2011).

The high population growth rates illustrated in Table 1 produce a high share of young dependents in the population in the long term, which we illustrate in Figure 1. The calculations are based on the same fertility and mortality schedules as in Table 1. The stable age structure of a population under a given r and l_x is given by Equation 4 (Euler-Lotka equation, in discrete form for 5 year age groups, see Slogett (2015), where l_x is the remaining life years at age x, and the numerator is the share of life years at time x, divided by the denominator, which is the sum of all life years in the population up to age ω. From the stable population equations, we can calculate the age distribution and show population pyramids by sex for our different scenarios. Using the age structure calculated from Equation 4, we also show various measures of demographic dependency.[10]

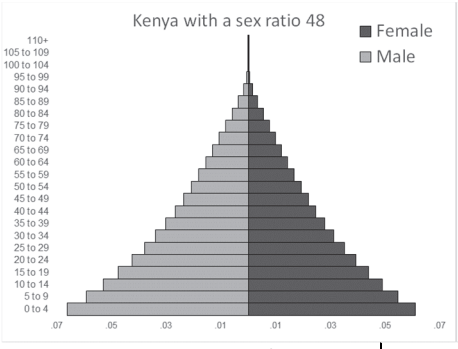
[10] We present three different measures. We calculated traditional dependency ratios for our population ((a0–14 + a65+) / a15–64). We also approximated how much of an adult working age individual's life (age 20–64) is spent on childcare in our stable population scenarios by summing the total share of life years in the 0–19 population and assuming that each life year of a child needs a third of an adult's life-year for education, child rearing, procreation, etc. The exact input a child needs will of course vary according to culture and context, and the value of ⅓ is only for illustration. We then divided the sum of time needed

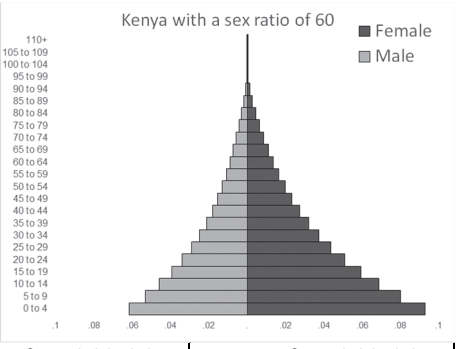
Figure 1. Population Pyramids at Equilibrium for Fertility Rates of the U.S. (2017) and Kenya (2014). Two different scenarios under two different sex ratios.





	Dependency ratio (population -14 and above 65 divided by population aged 15–64)	Ratio of available labor of primary age adults (age 20–64) spent on childcare	Ratio of available labor of primary age women (age 20–64) spent on childcare
U.S. 48	0.70	0.12	0.26
U.S. 60	0.66	0.19	0.27





	Dependency ratio (population -14 and above 65 divided by population aged 15–64)	Ratio of available labor of primary age adults (age 20–64) spent on childcare	Ratio of available labor of primary age women (age 20–64) spent on childcare
Kenya 48	0.66	0.16	0.27
Kenya 60	0.85	0.38	0.63

Note. The figure uses the same values and assumptions as in Table 1. It shows the eventual equilibrium age structure if fertility and mortality rates remain constant indefinitely.

to take care of the young population by the available time in the adult population to obtain a rough estimate of how much of all productive (and leisure) time of the population aged 20–64 (a0–19 ×⅓ / a20–64) must be spent on rearing the subsequent generations. We make the second calculations both for men and women and the denominator, and for only women in the denominator.

Equation 4:

$$a_x = \frac{l_x e^{r\,(x+2,5)}}{\sum_0^{\omega} l_x\, e^{r\,(x+2,5)}}$$

It is clear from Figure 1 that very high population growth rates (as seen in Table 1) cause a very young age structure. In the U.S., current fertility rates and sex ratios imply a shrinking population (from natural growth) with an old age structure, while similar fertility rates with a 60% female-biased sex ratio would instead result in a growing population with a younger age structure. Across the different populations shown in Figure 1, it is clear that female-biased sex ratios result in higher population growth, a higher share of women (naturally), and more resources that will have to be spent on supporting and rearing the young. For the U.S., with moderate/low fertility, the impact of age structure on different dependency ratios (see Figure 1) is relatively small and may even be beneficial. With Kenyan fertility levels, on the other hand, the high population growth and corresponding young populations with a highly biased female sex ratio would have consequences for the ability of adult members of society to adequately support the younger generations.

5. Discussion

5.1 Will Sex Selection Become Widespread in High-income Countries?

It should be noted that many people still express disapproval of non-medical parental sex selection, even with novel methods such as flow cytometry (Ethics Committee of the American Society for Reproductive Medicine, 2015). In a general population survey in the U.S. in 2006, only 18% of individuals aged 18–45 said they were positive, and 22% were undecided, if they had the option to use a cost-free, risk-free, non-invasive method to choose the gender of their child (Dahl et al., 2006). While most parents express a preference for a "balanced" family (i.e., at least one child of each sex), parents with one son are keener to do so than those with one daughter (Miranda et al., 2018).

On the contrary, we may have reasons to believe that the use of sex selection technologies will become more prevalent in the near future. According to a 2017 survey, 77% of fertility clinics in the U.S. that offer PGD also offer sex selection for non-medical reasons, which represents a substantial increase from 2006, when only 42% of clinics that offered PGD offered non-medical sex-selection (Capelouto et al.,

2018). Flow cytometry in combination with insemination is more affordable and less risky than PGD (which requires IVF) and less likely to be seen as morally objectionable, as it does not involve the discarding of fertilized eggs, meaning that access to this technology is likely to increase the use of sex selection in the general population.

Technologies associated with reproduction, including ART but also contraceptives of various kinds, have been highly controversial when introduced, and some remain so. However, we have consistently seen that attitudes toward different reproduction technologies have improved steadily over time. For example, IVF was once considered a highly divisive procedure. In the spring of 1972, the British magazine Nova ran a cover story suggesting that "test-tube babies" were "the biggest threat since the atom bomb" (Eschner, 2017; Henig, 2003). We can reasonably expect that at least part of current aversions to sex selection is due to a similar "yuck effect," which tends to dissipate as the use of the technology in question becomes normalized. Indeed, van Balen (2006) has described the technological trend toward more accessible, less invasive means of choosing the sex of a child as "nearly inevitable," and that any governmental countermeasures are likely to be largely ineffectual. We think it is plausible that the eventual prevalence of sex selection will be based primarily on the preferences among parents, rather than any technological barriers.

Some objections to sex selection concern some of the techniques used for this purpose. For example, those who object to abortions naturally also find their use as a means of sex selection objectionable. This "pro-life" stance sometimes also includes objections to IVF, especially when it involves the destruction of embryos. This makes the use of PGD for the purposes of sex selection an unattractive option. However, the use of flow cytometry does not involve killing a fetus or destroying an embryo and may therefore find less opposition among "pro-life" campaigners than other forms of sex-selection.

In sum, we argue that the combination of a latent and increasing daughter preference, new technology that facilitates sex selection (including flow cytometry), and increasing acceptance of ART in general suggest that sex selection is likely to become more prevalent in high-income countries over the coming decades.

5.2 Is Sex Selection a Self-reinforcing Practice?

In the study of cultural evolution, it has been noted that certain phenomena are self-reinforcing and increase in prevalence over time, while others are not. By contrast, other practices are self-limiting, in the sense that they produce outcomes that make them less prevalent or attractive. The phenomenon of sex selection may not only affect population growth directly (as demonstrated in section 4) but also have inter-

generational consequences over multiple generations. For most practices, it can be observed that children are more likely to resemble and copy the behavior of their parents than that of unrelated members of society (Bussey and Bandura 2004). If this also applies to norms and fertility practices, this will affect how prevalent the preference is in the next generation, as those parents who have fertility preferences that promote population growth will have more children, and those children will (often) share their parents' preferences (Kolk et al., 2014).

A practice can be self-reinforcing at both an individual and a group level (Murphy & Wang, 2001). For example, if individuals with a certain trait (e.g., a preference for having many children) have more children, and those children in turn also have that trait, the preference for having many children will increase in prevalence over time. Likewise, a group (e.g., a religious group or an ethnicity) where membership is inherited across generations will also increase in relative prevalence if its members have more children on average. In both cases, the practice will become more common in the population over time. In the context of this study, we argue that the practice of using sex selection technologies to select female children in high-income countries could become a self-reinforcing process, both at a population and a subpopulation level, in ways that the practice of selecting male children in some countries has not.

As shown above, populations with female-biased sex ratios have higher rates of population growth. As a population grows, the norms and practices of that population become more prevalent, all other things being equal. By contrast, selecting male children reduces population growth, and thus over time reduces the global impact of the norms of male-biased populations. Parents with an unusually strong preference for daughters may therefore decide to use sex selection, and then their daughters (and possibly sons) will themselves be more likely to sex-select than their peers.

This mechanism may be reinforced if social learning of practices and norms is itself sexually biased (Bandura & Walters, 1977; Bussey and Bandura 2004).[11] In other words, daughters may be more influenced in their reproductive choices by their mothers than by their fathers (Murphy, 1999). This means that women who more strongly prefer daughters are more likely to have offspring that will inherit this daughter preference. Assume that individuals in a population of women either want to select daughters or do not. Both will have the same number of offspring, but the women that select daughters are going to have more daughters and thus are more likely to transfer their preference for female sex selection to their offspring than the women that do not select daughters. Moreover, since our model assumes

[11] It has been noted that this sexual bias may explain how female infanticide can become an entrenched practice, even when it leads to fewer children and lower population growth (Strimling, Elrath & Richerson, unpublished).

that having more daughters leads to higher population growth, the women that select for daughters are going to have more grandchildren than women that do not select daughters. These grandchildren are also more likely to be daughters than the grandchildren of women that do not select for daughters, and they are more likely to inherit the preference for sex selection. In other words, the norms that are conducive of female sex selection are both adaptive in a demographic/natural selection perspective (a higher r) (since they produce a larger number of grandchildren) and create more "vessels" (i.e., daughters) for spreading those norms.

Moreover, since women on average have a stronger preference for daughters than men, in each generation daughter preferences (and selection for daughters) may become more common as women become a larger share of the population. This will be particularly true for the increasing share of women who choose to have children without a male co-parent, either as single mothers or in same-sex couples. If these groups both have more daughters than the average individual in a society, and their children share their preferences for sex selection, both the share of women raising children without men and the practice of sex selection may increase accordingly.

The cultural evolutionary logic above suggests that even if only a small minority of a population is positively dispositioned to sex selection, mechanisms exist through which this practice could become increasingly commonplace in each subsequent generation. If the self-reinforcing dynamic of this process proves to be correct, we should expect that populations with a female-biased ratio will be increasingly common in high-income countries. However, it is possible to make the case for a limit to an ever-increasing prevalence of sex selection, which we explore in the following section.

5.3 Will there be Counteracting Mechanisms to Sex Selection?

When assuming higher population growth among populations with female-biased sex ratios, we have used demographic calculations in which the availability of male partners is completely independent of the fertility rates of women in the population. This is common in demographic analysis, but such cases typically do not foresee very biased sex ratios. Similarly, in section 5.2, we highlighted a mechanism by which the female sex ratio would continue to increase. Both assumptions that (a) the sex ratio will not affect the age-specific fertility rates for a female in a population and (b) a cultural evolutionary mechanism will steadily increase the preference for female sex selection are almost certainly unrealistic for scenarios that deviate significantly from a balanced sex ratio. Indeed, at some level of bias in the sex ratio, it is reasonable to expect that other societal mechanisms may counteract these trends.

Below, we discuss such possibilities, beginning with arguments from evolutionary biology on why sex ratios tend to be balanced by natural selection and why this is not the case in our scenario with ART, followed by other mechanisms that nevertheless will also eventually balance the sex ratio at some level.

In species with sexual reproduction that are under natural selection, sex ratios are nearly close to 50–50 through self-correcting evolutionary processes. The evolutionary mechanisms according to which offspring of the sex that is temporarily underrepresented will have greater reproductive success is known as Fisher's principle (Hamilton, 1967). However, this logic that appears without few exceptions for natural selection is not applicable to the opportunity to procreate in our scenario, since ART removes the link between reproductive success and the sex ratio for women (i.e., as long as there is minimal number of men in the population, female reproduction is independent of the sex ratio). Nevertheless, the self-balancing mechanisms related to childrearing described in our first and second objections can be seen as analogous scenarios through which a sex ratio would be stabilized at an equilibrium, thereby stopping a process that would otherwise gradually increase the share of women.

A first counterargument is that while the biological/technological constraints of sex selection may be relaxed with the help of ART, as long as most childbearing takes place in different-sex partnerships, a deficit of males will constrain childrearing, availability of fathers, and eventual fertility rates. Traditional demographic models (including ours) assume that women are largely unconstrained by the availability of male partners for their fertility choices. This may be reasonable for sex ratios close to 50–50, but it becomes increasingly implausible with very unbalanced sex ratios, even if ART removes the biological necessity of males for female childbearing. We find it plausible that a male deficit would eventually make very unbalanced sex ratios unlikely, though we note that it does not apply to single mothers or to female same-sex relationships. Indeed, as we have noted, the same cultural processes that increase the female sex ratio may also increase the share of women that choose to procreate and rear children without men.

A second mechanism that may counteract a very large share of women in a population is how societies need to adjust the ways in which they provide resources for children. As we show in section 4.3, very rapid population growth leads to unbalanced dependency ratios between the young and adults. Similarly, if an increasing share of women choose to raise daughters by themselves, it seems likely that a single woman raising a child alone would settle for, on average, fewer children than what a couple would. Similarly, in a female same-sex relationship, the desired number of children per woman is very likely to be lower than that in a different-sex relationship; this is clear from the demography of same-sex parenthood shown by Kolk and Andersson (2020).

A third objection is that people may find a very unbalanced sex ratio "unnatural" or "disagreeable." More recently, feminist scholars have also objected to this practice. For example, Arianne Shahvisi (2018) argues that sex selection for the purposes of "family balancing" entrenches heteronormative stereotypes and misuses the moral mandate of reproductive autonomy. Elsewhere, Strange and Chadwick (2010) contend that prohibitive legislation against non-medical sex selection is justified because sex selection promotes restrictive conceptions of sex, gender, and family. Ultimately, if societies find an unbalanced sex ratio undesirable, they may adjust social policies to make such outcomes less likely. It is also plausible that parental preferences themselves may become increasingly less daughter-biased if we see a very biased female sex ratio. The preferences we see in high-income countries for a moderate daughter preference, may look quite different if the sex ratio is strongly biased towards females.

A fourth objection is that our model assumes that female preferences for the number of children will remain constant (for example, at two children per woman), regardless of the sex ratio. This seems unlikely, especially in female same-sex relationships and perhaps to a lesser extent in single mothers. However, it more likely that the average preference for the number of children will be reduced by less than the ratio of women in the population will increase. Assume an unrealistic scenario where the female share of births is 100%. To offset the increase in fertility in this scenario the average woman would have to reduce the number of children they prefer by 50%. Any reduction smaller than that would result in an increase in the population growth.

Taken together, the arguments above suggest that the potential cultural mechanisms that would increase the share of women in the population are more likely to have a moderate rather than substantial effect on the future equilibrium sex ratio, even if daughter preference becomes increasingly widespread, since counteracting mechanisms may limit the prevalence of sex selection for daughters. Based on the arguments above, we find it plausible that female-biased sex ratios will eventually reach equilibrium and that this equilibrium will not be particularly extreme. However, we do not know at which point this equilibrium will be reached.

6. Conclusions

The desire to select the sex of one's children is ancient in origin. In recent decades, selective abortions have mostly been used in low- and middle-income settings, and overwhelmingly to select male children. More recently, modern technology has made sex selection an inexpensive and non-invasive possibility that is less fraught with moral and medical concerns than abortion. We have argued that current trends

suggest that this technology will become more accepted and more widely used over time in high-income countries, where parents now seem to prefer female children over male children. Whether this turns out to be true is uncertain, but will depend on social trends and norms, the development of which is difficult to predict. Significantly, however, it would appear that governments can do little to restrict the use of flow cytometry, as doing so would involve legally unlikely infringements on bodily autonomy. We have also argued that if sex selection technology were to become routinely used to select female children, this practice may have a self-reinforcing dynamic, potentially leading to a consistent and durable bias in the sex ratio. In section 4, we described how such a sex ratio may affect population growth and the age structure, concluding that such effects are substantial and could help reach replacement rate fertility in high-income countries, while it would lead to rapid growth in countries with higher fertility.

The argument presented here is by its very nature speculative and based on the kind of uncertainty always associated with forecasting trends, but we argue that it also presents a plausible scenario. Our demographic calculations are not based on the empirical scenarios we consider most likely; rather, they aim to illustrate that the process will, over many generations, lead to substantial effects on demographic outcomes. While we do not foresee such demographic impacts to be substantial in the short term, over a longer time horizon their ramifications may be larger. If uptake of sex selection technology is small or moderate (which is plausible), the demographic effect may still be substantive. If used at lower frequencies, the dynamic effects that counteract a linear impact between sex ratios and population growth will also be less important, and a more linear relationship between more females and higher fertility will be observed. Ultimately, the aim of this article has not been to argue in favor or against the use of this technology, but to highlight its social impact over the long term. As our analysis makes clear, the consequences for population growth and social dynamics may be considerable over longer timescales.

References

Andersson, G., Hank, K., Rønsen, M., & Vikat, A. (2006). Gendering family composition: Sex preferences for children and childbearing behavior in the Nordic countries. *Demography*, 43, 255–267.

Arnold, F. (1997). *Gender preferences for children Demographic and Health Surveys Comparative Studies No. 23* Macro International Inc.

Arnold, F., & Zhaoxiang, L. (1992). Sex preference, fertility, and family planning in China. In D. L. Poston, Jr., & D. Yaukey (Eds.), *The population of modern China* (pp. 491–523). Springer.

Bandura, A., & Walters, R. H. (1977). *Social learning theory*. Prentice Hall.

Bennett, N. G. (1983). Sex selection of children: An overview. In N. G. Bennett (Ed.), *Sex selection of children* (pp. 1–12). Academic Press.

Bhatia, R. (2018). *Gender before birth: Sex selection in a transnational context*. University of Washington Press.

Blau, F. D., Kahn, L. M., Brummund, P., Cook, J., & Larson-Koester, M. (2019). Is there still son preference in the United States? (Working Paper No. 23816), *NBER Working Paper Series*. National Bureau of Economic Research.

Capelouto, S. M., Archer, S. R., Morris, J. R., Kawwass, J. F., & Hipp, H. S. (2018). Sex selection for non-medical indications: A survey of current pre-implantation genetic screening practices among US ART clinics. *Journal of Assisted Reproduction and Genetics*, 35, 409–416.

Dahl, E., Gupta, R. S., Beutel, M., Stoebel-Richter, Y., Brosig, B., Tinneberg, H. -R., & Jain, T. (2006). Preconception sex selection demand and preferences in the United States. *Fertility and Sterility*, 85, 468–473.

Drixler, F. (2013). *Mabiki: Infanticide and population growth in eastern Japan, 1660–1950*. University of California Press.

Dusenbery, M. (2020). What we don't know about IVF. *The New York Times*. 2020-04-16 https://www.nytimes.com/2020/04/16/parenting/fertility/ivf-long-term-effects.html

Eschner, K. (2017). In vitro fertilization was once as controversial as gene editing is today. *Smithsonian Magazine*. Retrieved from https://www.smithsonianmag.com/smart-news/vitro-fertilization-was-once-controversial-cloning-today-180964989/. Accessed November 20, 2020.

Ethics Committee of the American Society for Reproductive Medicine. (2015). Use of reproductive technology for sex selection for nonmedical reasons. *Fertility and Sterility*, 103, 1418–1422.

Fuse, K. (2013). Daughter preference in Japan: A reflection of gender role attitudes? *Demographic Research*, 28, 1021–1052.

Gartrell, N., Hamilton, J., Banks, A., Mosbacher, D., Reed, N., Sparks, C. H., Bishop, H. (1996). The national lesbian family study: 1. Interviews with prospective mothers. *The American Journal of Orthopsychiatry*, 66, 272–281.

Goldberg, A. E. (2009). Heterosexual, lesbian, and gay preadoptive parents' preferences about child gender. *Sex Roles*, 61, 55–71.

Guilmoto, C. Z. (2009). The sex ratio transition in Asia. *Population and Development Review*, 35, 519–549.

Guilmoto, C. Z. (2012). Son preference, sex selection, and kinship in Vietnam. *Population and Development Review*, 38, 31–54.

Guilmoto, C. Z., & Tove, J. (2015). The masculinization of births: Overview and current knowledge. *Population*, 70, 185–243.

Hamilton, W. D. (1967). Extraordinary sex ratios: A sex-ratio theory for sex linkage and inbreeding has new implications in cytogenetics and entomology. *Science*, 156, 477–488.

Hank, K. (2007). Parental gender preferences and reproductive behaviour: A review of the recent literature. *Journal of Biosocial Science*, 39, 759.

Hank, K., & Kohler, H. -P. (2000). Gender preferences for children in Europe: Empirical results from 17 FFS countries. *Demographic Research*, (2)1

Harper, J. C., & SenGupta, S. B. (2012). Preimplantation genetic diagnosis: State of the ART 2011. *Human Genetics*, 131, 175–186. https://doi.org/10.1007/s00439-011-1056-z.

Henig, R. M. (2003). Pandora's baby. *Scientific American*, 288, 62–67.

Henry, L. (1961). Some data on natural fertility. *Eugenics Quarterly*, 8, 81–91.

Higginson, T. M., & Aarssen, W. L. (2011). Gender bias in offspring preference: Sons still a higher priority, but only in men—women prefer daughters. *The Open Anthropology Journal,* 4 (1), pp 60-65. http://dx.doi.org/10.2174/1874912701104010060

Högbacka, R. (2008). The quest for a child of one's own: Parents, markets, and transnational adoption. *Journal of Comparative Family Studies*, 39, 311–330.

Human Fertility Collection—Kenya. (n.d.). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). Retrieved from https://www.fertilitydata.org. Accessed October 1, 2020.

Human Fertility Database—U.S. & Taiwan. (n.d.). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). Retrieved from https://www.humanfertility.org. Accessed October 1, 2020.

Human Mortality Database—U.S. (n.d.). Max Planck Institute for Demographic Research (Germany) and Department of Demography, UC Berkeley (U.S.). Retrieved from https://www.mortality.org. Accessed October 1, 2020.

Igbinedion, B. O.-E., & Akhigbe, T. O. (2012). The accuracy of 2D ultrasound prenatal sex determination. *Nigerian Medical Journal: Journal of the Nigeria Medical Association*, 53, 71–75. https://doi.org/10.4103/0300-1652.103545.

Jain, T., Missmer, S. A., Gupta, R. S., & Hornstein, M. D. (2005). Preimplantation sex selection demand and preferences in an infertility population. *Fertility and Sterility*, 83, 649–658.

James, W. H. (1987). The human sex ratio. Part 1: A review of the literature. *Human Biology*, 59, 721–752.

Johnson, S. D. (1994). Sex ratio and population stability. *Oikos*, 69, 172–176. https://doi.org/10.2307/3545299.

Karabinus, D. S., Marazzo, D. P., Stern, H. J., Potter, D. A., Opanga, C. I., Cole, M. L., Johnson, L. A., & Schulman, J. D. (2014). The effectiveness of flow cytometric sorting of human sperm (MicroSort) for influencing a child's sex. *Reproductive Biology and Endocrinology*, 12, 106. https://doi.org/10.1186/1477-7827-12-106.

O'Neill, K., Aghaeepour, N., Spidlen, J., Brinkman, R., (2013). Schematic diagram of a flow cytometer, from sheath focusing to data acquisition. *PLOS Computational Biology,* https://doi.org/10.1371/journal.pcbi.1003365

Kolk, M. (2011). Deliberate birth spacing in nineteenth century northern Sweden. *European Journal of Population*, 27, 337–359.

Kolk, M., & Andersson, G. (2020). Two decades of same-sex marriage in Sweden: A demographic account of developments in marriage, childbearing, and divorce. *Demography*, 57, 147–169.

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, 20132561. https://doi.org/10.1098/rspb.2013.2561.

Kolk, M., Schnettler, S. (2013). Parental status and gender preferences for children: Is differential fertility stopping consistent with the Trivers–Willard hypothesis? *Journal of Biosocial Science*, 45, 683–704.

Kupka, M. S., Ferraretti, A. P., De Mouzon, J., Erb, K., D'Hooghe, T., Castilla, J. A., Calhaz-Jorge, C., De Geyter, C., Goossens, V., & Strohmer, H. (2014). Assisted reproductive technology in Europe, 2010: Results generated from European registers by ESHRE. *Human Reproduction*, 29, 2099–2113.

Lamberts, R. W., Guo, D. P., Li, S., & Eisenberg, M. L. (2017). The relationship between offspring sex ratio and vasectomy utilization. *Urology*, 103, 112–116. https://doi.org/10.1016/j.urology.2016.11.039.

Lancaster, J. B., Kaplan, H. S., Hill, K., & Hurtado, A. M. (2000). The evolution of life history, intelligence and diet among chimpanzees and human foragers. In F. Tonneau, & N. S. Thompson, *Perspectives in ethology* (pp. 47–72). Springer.

Lee, R., & Mason, A. (2011). Lifecycles, support systems, and generational flows: Patterns and change. In R. Lee, & A. Mason, *Population aging and the generational economy: A global perspective* (pp. 79–106). Edward Elgar.

Leiblum, S. R., Palmer, M. G., & Spector, I. P. (1995). Non-traditional mothers: Single heterosexual/lesbian women and lesbian couples electing motherhood via donor insemination. *Journal of Psychosomatic Obstetrics and Gynecology*, 16, 11–20.

Leung, S. F. (1994). Will sex selection reduce fertility? Journal of Population Economics, 7, 379–392.

Lynch, R., Wasielewski, H., & Cronk, L. (2018). Sexual conflict and the Trivers-Willard hypothesis: Females prefer daughters and males prefer sons. *Scientific Reports,* 8, 15463. https://doi.org/10.1038/s41598-018-33650-1.

Mason, A., & Bennett, N. G. (1977). Sex selection with biased technologies and its effect on the population sex ratio. *Demography*, 14, 285–296.

MicroSort. (2020). Gender selection Mexico: Sex selection of your baby. *MicroSort International*. Retrieved from https://www.microsort.com/process/. Accessed November 20, 2020.

Miranda, V., Dahlberg, J., & Andersson, G. (2018). Parents' preferences for sex of children in Sweden: Attitudes and outcomes. *Population Research and Policy Review*, 37, 443–459.

Mungello, D. E. (2008). *Drowning girls in China: Female infanticide since 1650*. Rowman & Littlefield.

Murphy, M. (1999). Is the relationship between fertility of parents and children really weak? *Social Biology*, 46, 122–145.

Murphy, M., & Wang, D. (2001). Family-level continuities in childbearing in low-fertility societies. *European Journal of Population*, 17, 75–96. https://doi.org/10.1023/A:1010744314362.

Pollard, M. S., & Morgan, S. P. (2002). Emerging parental gender indifference? Sex composition of children and the third birth. *American Sociological Review*, 67, 600.

Reubinoff, B. E., & Schenker, J. G. (1996). New advances in sex preselection. *Fertility and Sterility*, 66, 343–350.

Sandström, G., & Vikström, L. (2015). Sex preference for children in German villages during the fertility transition. *Population Studies*, 69, 57–71.

Sen, A. (1990). More than 100 million women are missing. *The New York Review of Books*, 37, 61–66.

Sermon, K., Van Steirteghem, A., & Liebaers, I. (2004). Preimplantation genetic diagnosis. *The Lancet*, 363, 1633–1641. https://doi.org/10.1016/S0140-6736(04)16209-0.

Shahvisi, A. (2018). Engendering harm: A critique of sex selection for "family balancing." *Journal of Bioethical Inquiry*, 15, 123–137. https://doi.org/10.1007/s11673-017-9835-4.

Sharpe, J. C., & Evans, K. M. (2009). Advances in flow cytometry for sperm sexing. *Theriogenology*, 71, 4–10. https://doi.org/10.1016/j.theriogenology.2008.09.021.

Shi, L. (2017). *Choosing daughters: Family change in rural China*. Stanford University Press.

Sloggett A (2015). *Measuring fertility. In Population Analysis for Policy and Programmes* (103). Paris: International Union for the Scientific Study of Population. Available at http://papp.iussp.org/sessions/papp101_s04/PAPP101_s04_010_010.html. Accessed November 20, 2020.

Strange, H., & Chadwick, R. (2010). The ethics of nonmedical sex selection. *Health Care Analysis*, 18, 252–266. https://doi.org/10.1007/s10728-009-0135-y.

Tsuya, N. O., Wang, F., Alter, G., & Lee, J. Z. (2010). Prudence and pressure: Reproduction and human agency in Europe and Asia, 1700-1900. *The MIT Press*.

van Balen, F. (2006). Attitudes towards sex selection in the Western world. *Prenatal Diagnosis*, 26, 614–618.

Wacther, K (2014). *Essential Demographic Methods*, Harvard University Press.

Whittaker, A. M. (2011). Reproduction opportunists in the new global sex trade: PGD and non-medical sex selection. *Reproductive BioMedicine Online*, 23, 609–617. https://doi.org/10.1016/j.rbmo.2011.06.017.

Olle Torpman[1]

Inducement-Based Emissions Accounting[2]

In order to fairly divide the burdens of climate change, we need to know how much emissions belong to whom. For that purpose, we need a reliable emissions-accounting method. In this paper, I argue that none of the current emissions-accounting methods is satisfactory. I show this through a number of cases which all have intuitively clear answers to the question of who bears responsibility for emissions in these cases, but where the existing emissions accounting methods fail to provide these answers. I argue that this failure is due to the fact that none of them manages to identify the appropriate responsible-making feature of agents. Instead, I propose a new emissions accounting method – *inducement-based emissions accounting* – aimed at avoiding the problems faced by the current methods. I explain the role of this method and defend it against some objections.

[1] Institute for Futures Studies, Stockholm, olle.torpman@iffs.se

[2] Funding from Riksbankens Jubileumsfond (grant number M17-0372:1) is gratefully acknowledged. I would also like to thank Christian Barry, Stina Björkholm, Krister Bykvist, Tim Campbell, Göran Duus-Otterström, Hilary Greaves, Gustav Hedlund, Julia Mosquera, and Katie Steele for valuable comments.

1. Introduction

Assuming that countries, companies, or individuals have a duty to reduce their greenhouse gas emissions, say by 10% annually until reaching a net zero emissions target, we need to know which emissions belong to whom in order to know how much emissions each of them should reduce. And for that purpose, we need a reliable emissions-accounting method telling us who bears responsibility for which emissions.

In this paper, I argue that none of these emissions-accounting methods is satisfactory. I show this through a number of cases which all have intuitively clear answers to the question of who bears responsibility for emissions in these cases, but where the existing emissions accounting methods fail to provide these answers. I argue that this failure is due to the fact that none of them manages to identify the appropriate responsible-making feature of agents. Instead, I propose a new emissions accounting method – inducement-based emissions accounting – aimed at avoiding the problems faced by the current methods. I explain the role of this method and defend it against some objections.

The structure of the paper is as follows. In section 2, I clarify what an emissions accounting method is, and the context in which it is supposed to function. In section 3, I explain how and why the debate over emissions accounting methods is relevant. In section 4, I clarify how emissions accounting methods can be evaluated, and show why none of the existing emissions accounting methods is satisfactory. Section 5 introduces the inducement-based emissions accounting method. Sections 6-7 answer potential objections to this method. Section 8 concludes.

2. What Is an Emissions Accounting Method?

Several emissions-accounting methods have been proposed in the literature, providing different answers as to how much emissions each country or individual is responsible for (see, e.g., Skeie et al 2017). The main rivals discussed in the debate are:

> ***Extraction-Based Accounting*** **(EA)**: Responsibility for emissions is assigned to the *extractors* of the emissions-generating resources.
>
> ***Production-Based Accounting*** **(PA)**: Responsibility for emissions is assigned to the *producers* of emissions-generating goods and services.
>
> ***Consumption-Based Accounting*** **(CA)**: Responsibility for emissions is assigned to the final *consumers* of those goods and services.

> ***Mixed Accounting* (MA)**: Responsibility for emissions is instead shared between extractors, producers and consumers.

To clarify, EA assigns responsibility for emissions to those who mine iron ore, extract oil, coal, and gas from the ground, cut down forests, and so on (see, e.g., Steininger et al 2014: 75). On PA, responsibility is instead assigned to those who produce military weapons of the iron, burn the coal to produce electricity, refine the oil into fuels for cars and airplanes, and so on (e.g., Tukker et al 2020: 54). On CA, responsibility for emissions is assigned to those who use the military weapons, buy the electricity to heat their homes, drive the cars, and so on (e.g., Peters & Hertwich 2008; Duus-Otterström & Hjortén 2018). On MA, responsibility for emissions is divided between the different extractors, producers and consumers that are involved in the situation, depending on their respective roles (e.g., Lenzen 2007; Steininger et al 2016).

These emissions-accounting methods differ from one another as they identify different *responsible-making features*, as it were, of agents. A responsible-making feature (or set of features) is a (set of) feature(s) in virtue of which an agent is responsible for a certain action or outcome. What is common to all major emissions accounting methods, however, is that they take the crucial feature in this respect to be a causal connection (of some sort) between the agent and the emissions at issue: On EA it is the feature of being the initial extractor of the emissions-generating resources; on PA it is the feature of being the producer who refines these resources into emissions-generating goods and services; on CA it is the feature of being the final consumer of these emissions-generating goods and services.

It is not entirely clear from the literature, however, what *kind* of responsibility the emissions accounting methods are supposed to assign. At a minimum, they assign *attributive* responsibility in the sense that they come with a conceptual claim about who counts as an "emitter". In other words, they merely say which emissions belong to whom and why. On a more substantial interpretation, they also come with a normative claim about who bears *remedial* responsibility for emissions, thus telling us who should rectify or pay the costs for emissions (see, e.g., Steininger et al 2016; Tukker et al 2020; Torpman 2022). In this paper, I will stick to the minimal view, and thus take emissions-accounting methods to assign merely attributive responsibility for emissions.

On any of these views, however, emissions-accounting methods are not sufficient for assigning remedial responsibility for emissions. One reason for this is that, based on the "ought" implies "can" principle, there is a widely accepted *ability condition*. This condition implies that even if an agent is assigned responsibility for

emissions by the correct emissions accounting method, she is remedially responsible for them only if she is able to remedy them.[3]

Emissions-accounting methods are not even necessary for assigning remedial responsibility for emissions. The reason is that someone can be remedially responsible for emissions even if she does not count as an emitter on any emissions accounting method. Take the following example for clarification. Suppose that a person has given rise to zero net emissions but is so rich that she could easily remedy others' emissions without any personal sacrifice. Assume furthermore that these other people would have to make a substantial personal sacrifice to remedy their emissions without the help of the rich person. In such a case, it seems plausible that the rich person should remedy at least some of these emissions – even though she bears no attributive responsibility for them. This suggests that something should be said about the relevance of emissions accounting methods.

3. The Relevance of Emissions Accounting Methods

The example above puts into question the relevance of the debate over emissions accounting methods. For one reason, it puts into question one of the assumptions that underlies the debate: The assumption that the emitter should pay the costs for her emissions. In essence, this assumption is captured by:

> ***The Polluter-Pays Principle*** **(PPP)**: Those who have emitted should pay the costs for climate change, in proportion to their emissions.

The intuition that underlies PPP is supported by general considerations of fairness, according to which remedial responsibility for an outcome falls on those who contribute to that outcome. Perhaps, therefore, it is not surprising that PPP is popular among climate ethicists as well as policymakers. Note, moreover, that it is precisely due to PPP that the question under consideration (attempted to be answered by the emissions accounting methods) has been raised: who *is* an emitter?

PPP is not without rivals. For one reason, it fails to deal with cases as the one above, involving rich non-emitters. More generally, PPP has been criticized for disregarding such relevant factors as people's different abilities, and the different extents to which people have benefitted from the activities related to the emissions.

[3] There are several ways in which emissions could be remedied: (a) via mitigation measures, e.g., through absorption of greenhouse gas from the atmosphere via tree-plantation; (b) via adaptation measures, e.g., through constructions of sea walls or installations air-conditioners aimed at tackling effects from unmitigated climate change; or (c) via compensation measures, e.g., through direct payments to those who are harmed due to a failure to adapt to unmitigated climate change (e.g., Page 2016: 84).

The major rivals to PPP are:

> ***The Ability-to-Pay Principle*** **(APP)**: Those who are able to pay should pay the cost for climate change, in proportion to their ability.
>
> ***The Beneficiary-Pays Principle*** **(BPP)**: Those who have benefitted from climate change (or the actions leading to it) should pay the costs for climate change, in proportion to their benefits.

As is clear, none of these rival principles to PPP attribute responsibility for emissions to emitters *qua* emitters. Hence, none of them makes use of any emissions accounting method. In order to motivate the further discussion about the emissions accounting methods, it seems that we need to justify PPP over these rivals.

It becomes clear that something is lacking in PPP once we consider the intuitions that motivate APP and BPP. Take the intuition behind APP first. One might argue that the ability condition restricts the cases to which PPP and its accompanying emissions accounting methods are supposed to be applicable (see, e.g., Page 2011). But this does not help us all the way. The reason is that ability is not a binary concept, but rather a matter of degree. One agent might be *more* able than another, yet both are able. Suppose that agents A and B are both able emitters, where A has emitted more and where B is more able. Irrespective of how little more A has emitted than B, and of how much more able B is than A, PPP would assign more remedial responsibility to A than B. This is counterintuitive. If A has emitted just *slightly* more than B, and if B is *much* more able than A, the intuitive answer is that more remedial responsibility should be assigned to B than to A. This is not what PPP recommends.

The intuition behind BPP, on the other hand, says that people's different climate change-related benefits are relevant for their remedial responsibilities. To exemplify why this intuition is not explained by PPP, suppose that A and B have emitted equal amounts and that they are equally able to pay, but that only A has benefitted from the emissions and their climate effects. In this case, PPP (and APP) would divide responsibility equally between A and B, although the intuitive verdict assigns more responsibility to A in virtue of the fact that A, but not B, has benefitted from the emissions. BPP, however, would provide this answer.

Even if we want to add 'being able' or 'being a beneficiary of emissions' as responsibility-making features, our reasons remain for thinking that 'being an emitter' is such a feature. The reason is that we, in any case, want to be able to say that emitters bear more responsibility for emissions than non-emitters, other things being equal. To exemplify, suppose that A and B are equally capable of remedying emissions, and are equal beneficiaries of emissions, but that A has emitted more than B. In such a

case, A has a stronger duty to remedy emissions than B, other things being equal (see, e.g., Couto 2018).

This suggests that we nevertheless need to determine who is an emitter, and to what extent they emit, in order to give a full account of how remedial responsibility for emissions should be divided between people. And for that reason, an emissions accounting method is needed.

Still, one could question the role of emissions accounting methods in the overall climate ethics framework. Suppose we had an optimal carbon tax, so that climate externalities were internalized, and that the tax revenues were then used to pay for climate change mitigation. In this case, it seems to make little difference whether the tax is levied at the point of extraction, production, or consumption. It will, in whichever case, disincentivize carbon-intensive activities to the same extent: If the tax is levied at the point of consumption, fewer consumers will buy, so demand will be lower, and hence there will be less production and less extraction; if it is levied at the point of production, the associated products will end up more expensive for consumers and demand will drop in response; if it is levied at the point of extraction, it will make production more expensive, and thus make end products more expensive for consumers, which in the end will incentivize consumers to consume less. Given this, one might question what role emissions accounting methods play in our efforts to fight climate change.[4]

It is true that under ideal circumstances, where a carbon tax system has been successfully adopted and where all costs for emissions are thus internalized, there would be no point of assigning attributive responsibility for emissions. Indeed, under a carbon tax system, it would make little sense to debate whether polluters, the able, or beneficiaries should pay the costs for emissions. The simple reason is that there would then be no leftover, unpaid, social costs of emissions to distribute. Until such a carbon tax system is adopted, however, there are such unpaid costs the distribution of which it makes sense to debate. And it makes sense, under current nonideal circumstances, to say that emitters bear some responsibility in this respect, and that they should thus pay at least some of these costs (along the lines of PPP). Therefore, it also makes sense in current circumstances to debate who should count as an emitter in this regard – which is exactly what the debate on emissions accounting methods is about.

Of course, this does not mean that we should not work towards the implementation of a carbon tax system even under present nonideal circumstances. But doing so will not solve the problem of how to divide responsibility for emissions that occur

[4] I thank Hilary Greaves and Christian Barry for raising this worry.

until that implementation is completed. And, for that task, emissions accounting methods have an important role to play.

4. Why Existing Emissions-Accounting Methods Fail

Having established the role of emissions accounting methods, I will now move on to the question of how they can be evaluated. Typically, issues regarding causal responsibility, effectiveness, political and technological feasibility are considered relevant for the evaluation of emissions accounting methods. Roughly, these considerations can be boiled down into two main desiderata, that any satisfactory emissions-accounting method must meet (c.f., Steininger et al 2016; Mittiga 2019; Duus-Otterström 2022):

> *Fairness*: It provides an identification of the responsible agents, and explains *why* the identified agents are responsible.
>
> *Effectiveness*: It provides a feasible recommendation that contributes to solving issues of climate justice.

It is not unambiguous what the fairness desideratum requires. I will stick to the idea that an assignment of responsibility is fair if it is sensitive to people's different contributions and capacities (Steininger et al 2014). When it comes to the effectiveness desideratum it suffices for the argument in this paper to say that it implies technological and political feasibility, since a recommendation cannot be effective without being feasible in these respects.

In the remainder of this section, I will evaluate the existing emissions accounting methods on the basis of the fairness desideratum. I assume that the way to test whether such a method meets this desideratum is to see whether it can offer an intuitive answer to the question of how responsibility for emissions should be assigned in clear cases. I follow the existing literature when saying that goods or services "embody" emissions, by which it is meant that things come with, involve, or are attached to emissions (see Peters & Hertwich 2006; Davis & Caldeira 2010; Duus-Otterström & Hjorthen 2018; Tukker et al 2020).

To see why EA, firstly, fails to provide the intuitive answer to whom responsibility for emissions should be assigned, consider the following case:

> *New Invention*: A high-tech company invents a new smart device for which a certain rare earth mineral and certain amounts of fossil fuels, is needed.

> To be able to produce the device, the company pays an extractor to extract these minerals, giving rise to a large amount of emissions. Assume that no other agent is relevantly involved in the situation, and that without the company's influence the extractor would not have extracted the resources. To whom do these emissions belong?

In this case, the intuitive answer is that the emissions should be attributed to the producer – and not the extractor. Indeed, the producer's invention and then production of the new smart device is what initiates the process that leads to the emissions. However, this answer cannot be provided by EA. Hence, we might want to adopt PA instead, on which the producer would be responsible. But now, consider:

> *Current Demand*: A number of consumers demand certain goods that are not yet available on the market. This incentivizes a company to produce the goods demanded by the consumers. These goods yield carbon emissions that would never have been generated without the consumers' demand. Assume that no other agent is relevantly involved in the situation. To whom do these emissions belong?

Since the consumers' demand is what initiates the causal chain leading to the emissions in this case, the intuitive answer is here that the consumers bear responsibility for the emissions. However, this answer can be provided neither by EA nor by PA, but by CA. One might hence think that both EA and PA are unsatisfactory, and that CA should be adopted instead. However, CA fails in other clear cases. Consider a case similar to *New Invention*:

> *Market Introduction*: A high-tech company produces, manufactures and makes advertisement for a new smart device, which makes consumers buy and use it. There was no demand for the device before the producer's activities. This device embodies a large amount of emissions. Assume that no other agent is relevantly involved in the situation. To whom do these emissions belong?

In *Market Introduction,* the intuitive answer attributes the emissions to the company *qua* producer, simply for the reason that the producer triggers the process leading to the emissions. This verdict is implied by PA, but not by CA.

Altogether, these simple cases might suggest that the responsible part is *sometimes* the extractor, *sometimes* the producer, and sometimes the consumer. This might be seen as a motive to move towards an emissions accounting method such as

MA. But even if such a method would manage to yield the right answers in all above cases, it would fail in other clear cases. For instance, consider:

> *Defensive War*: Iron ore is extracted in Sweden and exported to Poland where it is used for production of military weapons. These military weapons are then used by Ukraine in self-defense against a Russian invasion. These weapons embody a large amount of emissions. Assume that no other agent is relevantly involved in the situation. To whom do these emissions belong?

In this case, it seems clear that the emissions should be attributed to Russia. However, this cannot be explained by MA – since Russia is neither of an extractor, a producer, or a consumer in this case. As this moreover implies, it cannot be explained by any of EA, PA or CA either. Consequently, this shows that there are clear cases where all standard emissions-accounting methods fail to correctly answer who is attributively responsible for emissions.

What conclusion should we draw from this? One would be that the existing emissions accounting principles fail to identify the correct responsible-making feature. This means that, in cases where the right verdict holds the extractor or the producer or the consumer attributively responsible for the emissions, the explanation is not that this is because the agent is an extractor, a producer, or a consumer per se. Instead, the right explanation would be that the responsible agents in these cases possess some common feature that is independent of them being extractors, producers, consumers – or invaders.

5. Inducement-Based Emissions Accounting (IA)

What is the most crucial feature of all responsible parts in the cases above? The apparent answer to this question seems to be that they are all inducers of the emissions in those cases. In *New Invention*, the producer is responsible in virtue of inducing the extraction and production which give rise to the emissions; in *Consumer Demand*, the consumers are responsible in virtue of inducing the emissions-generating activities undertaken by the producer; in *Market Introduction*, the producer is responsible in virtue of inducing the production as well as the consumption of the emissions-generating device; and in *Defensive War*, Russia is responsible because of inducing Ukraine's defensive weapon-use that yields the emissions.

This suggests that we should adopt the following emissions accounting method:

> ***Inducement-Based Accounting*** **(IA)**: Responsibility for emissions is assigned to agents in proportion to their respective inducements of emissions.

Although one might have an intuitive grasp of what "inducement" means, it is not clear when considered more carefully. In any case, it should not be understood in counterfactual terms, since it is true of most agents involved in the cases above that the emissions would not have occurred if they had not acted as they did. Instead, "inducement" should be understood in probabilistic terms, where an agent counts as an inducer of emissions if they saliently increase the probability of those emissions. I say "saliently" increase the probability of the emissions, by which I mean making a significant contribution to its occurrence, simply in order to rule out non-salient factors (i.e., background conditions) that might also increase the probability of the outcome. On this account, then, an agent is the main responsible part for an outcome if their action increases the probability of the outcome more than the actions of other agents. If an agent does not at all increase the probability of a certain outcome, then she bears no attributive responsibility for that outcome.

If I am right about this, the existing emissions accounting methods are thus mistaken about what is the appropriate responsible-making feature of agents. While the cases above show that it is (often) appropriate to assign responsibility to either extractors or producers or consumers, this should be considered an *implication* of a method like IA – rather than being considered an emissions accounting method in itself.

Of course, the arguments above are too quick to prove that IA is more plausible than its rivals. First of all, there might be alternative explanations as to why the existing views fail to provide the correct answers in the considered cases.

For instance, one might want to argue that by unjustly invading Ukraine, Russia is committing a severe wrongdoing in *Defensive War*, which in turn triggers a duty to correct for this wrongdoing, and that this corrective duty is what explains why Russia is responsible in this case. But this corrective duty is victim-regarding, in the sense that it is a duty to Ukraine considered as a victim of Russia's invasion. It is thus unrelated to the duty to remedy the involved emissions, and hence cannot explain why Russia is responsible for the related emissions.

Still, as I argued in section 3, there are (plausibly) other responsible-making features of agents besides that of 'being an emitter'. One could thus argue that what explains our intuitions in *Defensive War* is that we find Russia responsible for emissions not in virtue of being an emitter in this case, but in virtue of possessing some other such responsible-making feature. At closer scrutiny, however, it seems that none of the features that would be potentially relevant in the case of emissions, such as 'being able' or 'being a beneficiary of emissions', are relevant here. It is not

because we find Russia a beneficiary of emissions, or particularly able to remedy them, that we find them responsible for them. It is rather because we find Russia to be a contributor to emissions in this case.

On that note, however, it seems that Russia is only inducing the emissions that are due to the *use* of the weapons, and not the emissions that are due to the production of the weapons, or the *extraction* of the resources used in this production. This would mean that even IA fails to assign responsibility for these emissions to Russia in this case. At a closer look, however, some of the related extractions and productions are in fact induced by threats from Russia, implying that IA would assign responsibility to Russia for a corresponding proportion of these emissions. Still, the remaining emissions should be attributed to whichever part – Sweden, Poland, or Ukraine – is the inducer of these emissions. This, however, should not be seen as a shortcoming of IA but rather as a positive feature of IA, as it opens for the possibility of assigning responsibility to several different agents in one and the same situation, depending on the extent to which they induce emissions in that situation.

6. Potential Fairness-Objections to IA

Despite the abovementioned advantages of IA over its rivals, it is still not clear that IA manages to satisfy the fairness desideratum. It might fail to yield the correct answer in independent cases. For instance, consider:

> *The Contract:* A consumer decides to reduce her net emissions. She finds out about a company offering to offset her emissions in an appropriate way, and contracts them to do so. However, the company violates the contract, takes the money but refuses to conduct any offsetting. This leads to more emissions. Who is responsible for these emissions?

In this case, it seems obvious that the company is the main responsible part for these emissions. However, it appears as if the company has not induced any emissions at all, since the emissions are already made by the consumer. Hence, it seems that the intuitive answer in this case cannot be provided by IA.

Here, however, it should be emphasized that what matters fundamentally is *net* emissions. Hence, it does not suffice to look only at the activities yielding the initial *positive* emissions induced by the consumer, since we need also look at any related *negative* emissions. Since the consumer would in fact (we may assume) have given rise to zero *net* emissions had the company kept to the contract and offset her emissions appropriately, the company is in fact an inducer of net emissions once they fail to do so. Hence, IA can nevertheless provide the intuitive answer in this case.

Talking of the possibility of yielding net zero emissions, another case that might appear problematic for IA is:

> *Climate Neutrality*: The world has become climate neutral, and hence there are no longer any inducers of net emissions around. Still, climate change affects people, since enormous amounts of greenhouse gases are already in the atmosphere. Hence, there is a need for measures of adaptation and compensation. So, who should pay for *that*?[5]

In this case, IA would not attribute responsibility for emissions to any living agent, since no one is anymore inducing any net emissions. And this may be counter-intuitive, since we want someone to pay for the mitigation and adaptation measures needed.

Here, it is first worth noting that *if* this case poses a problem for IA, then it would pose a problem for any emissions accounting method. In a climate neutral world, where there would be no net emissions, there would be neither any extractors of emissions-generating resources, nor any producers or consumers of emissions-generating goods or services. Consequently, IA would not fare worse than these other emissions accounting methods in *this* respect.[6]

More importantly, however, we have already seen that emissions accounting methods are not necessary for remedial responsibility for emissions. Indeed, such methods make sense only in cases which involve emitters. But in *Climate Neutrality*, there are no emitters. Hence, any remedial responsibility should in this case be divided based on some *other* consideration – perhaps related to people's different abilities or benefits from emissions. Once again, this suggests that being an emitter is not the one and only responsible-making feature in the context of climate change.

7. Potential Effectiveness-Objections to IA

Even if IA would win over its rivals from the perspective of fairness, it is less clear that it would win from the perspective of effectiveness. For one reason, it is not clear that IA would be of any help when determining the extent of an agent's responsibility for emissions, at least in comparison with other emissions accounting methods. It appears to be easy to identify the initial extractor of emissions-generating re-

[5] See Mittiga (2019: 185-6). This relates to the problem of historical emissions. See Page (2008: 559), Duus-Otterström (2014), and Torpman (2022).

[6] This is a bit too simplified, however. In fact, CA can capture some hitorical emissions to the extent present people consume goods produced in the past. See Torpman (2022). However, this does not affect my present argument.

sources. Likewise, it is relatively easy to identify the producers who refine these resources into goods and services, as well as the final consumers of these products. However, it appears to be much harder to identify all inducers of emissions, and to determine the respective extents to which they induce emissions.

A similar worry is raised by Steininger et al (2014: 78), saying that "...we would have to be able to ascribe [...] relative shares of contribution, and thus relative shares of causation. Unfortunately, this is impossible...". They continue: "[T]he only robust statement that we can make is that consumers and producers both contribute emissions, but it is hardly possible to justify a statement to the effect that one of them contributes more than the other or that they both contribute equally much" (2014: 78). In Steininger et al (2016: 1), they explain the complexity further by saying that "[f]irst, there is the problem of identifying those agents who are causally responsible for the harmful emissions. [...] Second, there is no agreement on how to determine relative causal shares in any instance of joint causation of harm, where each single agent's contributions were neither necessary nor sufficient". This point is rehearsed by Mittiga (2019: 173), saying that "[p]arsing the causal impacts of consumers versus producers [...] is – philosophically and practically – infeasible" (see also Tukker et al 2020: 54-55).

There is certainly something into these worries. But there are reasons not to exaggerate them and their implications for the evaluation of IA. Although it is hard to make practical use of IA in many cases, it is not so in all cases. This is shown by the cases brought up in section 3 above. This suggests that IA can be used to assign responsibility for emissions *in those cases*.

For example, existing life cycle analyses provide detailed information about the origin and extent of emissions in the supply chain of goods and services, from which it can be inferred how an agent can lessen inducement of emissions and hence responsibility for emissions in a broad range of cases. From an extractor's perspective, it is possible to lessen inducement of emissions and hence responsibility for emissions by reducing extraction, for instance by recycling already extracted resources. From a producer's perspective, it is possible to reduce responsibility for emissions by producing more sustainable and long-lasting products, and by using renewable energy and recycled or carbon-free resources. From a consumer's perspective, it is possible to lessen one's degree of responsibility for emissions by consuming less, or more climate friendly, for instance by consuming second-hand or recyclable/reusable and long-lasting goods and services (Tukker et al 2020: 57).

In cases where it is harder to exactly determine an agent's degree of inducement of emissions, we could infer some rule of thumb for responsibility assignments that approximates IA. There are several possibilities in this regard. For instance, responsibility could be assigned to the *main* inducer of emissions. In this respect, one of the

existing emissions accounting methods, such as PA or CA, could be used. This is what Mittiga argues, when saying that "[c]hoosing between [PA] and [CA] is less a matter of which method better captures contribution (which both do imperfectly), but of which performs better with respect to pertinent ethical factors: especially, fairness, environmental efficacy, and cost-effectiveness" (2019: 173). He moreover goes on to argue that "[CA] is superior in these three regards…".

Although I agree with Mittiga that CA might be superior to PA in many regards, I do not think that CA should be used singularly and universally. As shown in section 3, CA fails to yield the correct answer in several types of cases. Likewise, PA and EA fail in other types of cases, for which reasons they should neither be adopted singularly and universally. As *Defensive War* shows, even MA fails to capture the relevant inducers of emissions in certain important cases. The more obvious answer, therefore, is that IA should be used to determine *which* of these methods to be used as a rule for assigning responsibility for emissions in particular cases. This implies that extractors will sometimes be held responsible, while producers and consumers will be so held at other times.

The problem is, then, what we should do in cases where IA cannot be used to determine which of these rules is most plausible to apply, or in cases where we know that neither extractors, producers nor consumers are the main responsible part (such as in *Defensive War*). Here it seems that we would do best by following a 'principle of insufficient reason' and thus divide responsibility evenly between all relevantly involved agents (where "relevantly involved" means "being an inducer of emissions"). This idea would suggest a revised version of MA, call it MA*:

> ***Revised Mixed Accounting* (MA*)**: Responsibility for emissions in a certain situation is shared evenly between all involved inducers of emissions in that situation.

Perhaps, one might wonder why we should not then move directly to MA* instead of going via IA. The answer is twofold. First, MA* applies only to cases where different agents' inducements of emissions cannot be determined. As we have seen above, it is not generally the case that we lack the knowledge needed for IA to yield determinate recommendations of responsibility assignments. And in those cases, IA should be applied directly. Second, MA* fails to meet the fairness desideratum, since it is uncapable of explaining *why* all involved agents should share equal responsibility for the involved emissions. However, if considered as a mere rule of thumb supposed to supplement IA, MA* does not have to provide such an explanation. That explanation would then be provided by IA.

It should finally be noted that it is one thing to answer the theoretical question

of which feature makes someone responsible for something, and quite another thing to answer the practical question of how to identify the agents who possess this feature. Just because IA might be hard to apply in some cases does not mean that it should be rejected. It all depends on how weighty the effectiveness desideratum is compared to the fairness desideratum, and the extent to which a principle like IA can be made further practicable in order to meet this desideratum. This is not a task ultimately for climate ethicists, but rather for economists, political scientists, and other social scientists to conduct together with policymakers.

8. Conclusion

In this paper, I have argued that the current emissions-accounting methods are implausible for failing to identify the responsible-making feature of agents. Hence, they also fail to meet the fairness desideratum. I have thus proposed an inducement-based emissions accounting method which manages to identify the responsible-making feature, and thus meets this desideratum. While this method has some troubles meeting the effectiveness desideratum in certain cases, I argue that it should be used to determine which of existing emissions accounting methods to use for assigning responsibility for emissions in those cases. This suggests that the existing emissions accounting methods have a mere derivative role to play, and that none of them should be used singularly or universally.

References

Couto, Alexandra (2018), "The Beneficiary Pays Principle and Strict Liability: exploring the normative significance of causal relations". *Philosophical Studies* 175, 2169–2189 (2018). https://doi.org/10.1007/s11098-017-0953-y

Davis S. and Caldeira, K., 2010. "Consumption-based accounting of CO2 emissions". *PNAS*, 107 (12), 5687–5692.

Duus-Otterström, Göran (2014), "The problem of past emissions and intergenerational debts", *Critical Review of International Social and Political Philosophy*, 17:4, 448–469, DOI: 10.1080/13698230.2013.810395

Duus-Otterström, Göran & Fredrik D. Hjorthen (2018), "Consumption-based emissions accounting: the normative debate", in *Environmental Politics*, Vol 28, No 5.

Duus-Otterström, Göran (2022), "Sovereign States in the Greenhouse: Does Jurisdiction Speak Against Consumption-Based Emissions Accounting?", *Ethics, Policy & Environment*, DOI: 10.1080/21550085.2022.2061253

Lenzen, M., et al., 2007. "Shared producer and consumer responsibility – theory and practice". *Ecological Economics*, 61 (1), 27–42.

Mittiga, Ross (2018), "Allocating the Burdens of Climate Action: Consumption-based Carbon Accounting and the Polluter-pays Principle", in B. Edmondson and S. Levy (eds.), *Transformative Climates and Accountable Governance*, Palgrave Studies in Environmental Transformation.

Page, A. Edward (2008), "Distributing the burdens of climate change", *Environmental Politics*, 17:4, 556-575, DOI: 10.1080/09644010802193419

Page, A. Edward (2011), "Climatic Justice and the Fair Distribution of Atmospheric Burdens: A Conjunctive Account", *The Monist*, Volume 94, Issue 3, July 2011, pp. 412-432, https://doi.org/10.5840/monist201194321

Peters, G. P. and Hertwich, E. G. (2006), "Pollution embodied in trade: the Norwegian case", *Global Environmental Change*, 16 (4), 379–387. doi:10.1016/j.gloenvcha.2006.03.001

Peters, Glen P. & Hertwich, Edgar G. (2008), "Post-Kyoto greenhouse gas inventories: production versus consumption", *Climate Change* (2008) 86:51–66.

Skeie, Ragnhild B., et al (2017), "Perspective has a strong effect on the calculation of historical contributions to global warming", *Environmental Research Letters*, 12 024022

Steininger, K., Lininger, C., Droege, S., Roser, D., Tomlinson, L., & Meyer, L. (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, Karl W., Lininger, Christian, Meyer, Lukas H., Muñoz, Pablo and Schinko, Thomas (2016), "Multiple carbon accounting to support just and effective climate policies", *Nature Climate Change*, Vol 6, January 2016. DOI: 10.1038/NCLIMATE2867

Torpman (2022), "Consumption-Based Emissions Accounting and Historical Emissions", Ethics, *Policy & Environment*.

Tukker, A., Hector Pollitt & Maurits Henkemans (2020), "Consumption-based carbon accounting: sense and sensibility", *Climate Policy*, 20:sup1, S1-S13, DOI: 10.1080/14693062.2020.1728208

Stephen M. Gardiner[1]

The Ethical Risks of an Intergenerational World Climate Bank (as Opposed to a Climate Justice World Bank)[2]

Recently, John Broome and others have been arguing that future generations should shoulder the burden of climate mitigation, through a strategy sometimes called "making the grandchildren pay". This strategy appeals primarily to the idea of making a Pareto improvement that delivers "efficiency without sacrifice", particularly for the current generation. More generally, "making the grandchildren pay" is said to have major pragmatic advantages, since it appeals to self-interest rather than morality. Broome argues that economists should take up the task of implementing the strategy by designing a new institution along the lines of the World Bank and International Monetary Fund. He dubs this institution "the World Climate Bank". In this paper, I argue against Broome's proposal on both ethical and pragmatic grounds. Instead of a world climate bank that shifts all the burdens of climate action to the future, what is needed is a Climate Justice World Bank that respects principles of global and intergenerational ethics.

[1] Professor of philosophy at University of Washington, e-mail: smgard@uw.edu

[2] The first version of this paper was written in 2019-2020. One section of Gardiner 2021a summarizes part of that version. I am grateful to participants at the climate futures workshop at Princeton University and to audiences at Concordia University, the International Society for Environmental Ethics, and the University of Graz. I especially thank Gustaf Arrhenius, Alyssa Bernstein, Tim Campbell, Phil Carafo, Matthias Fritsch, Aaron James, Lukas Meyer, Matthew Rendall, and Olle Torpman.

Some climate ethicists propose that climate mitigation should be funded through a general strategy they dub "making the grandchildren pay" and "efficiency without sacrifice" (e.g., Rendall 2011; Broome 2012; Maltais 2015). The central objective of this strategy is to pass on to future generations the full costs of the technological transition away from fossil fuels, in order to avoid imposing any sacrifice on the current generation and so overcome political inertia. To implement the strategy, John Broome and Duncan Foley have called for the creation of a new financial institution that they name 'the World Climate Bank' ('WCB'). This institution would resemble the International Monetary Fund and the conventional World Bank. According to Broome, economists should take on the responsibility of designing a global financial institution which has the specific aim of shifting all of the burdens of climate mitigation onto the future (e.g., Broome 2012; cf. Rendall 2021).

The proposal to "make the grandchildren pay" is an interesting one, and borne out of a genuine desire to aid future generations in confronting the climate threat. Its proponents deserve respect for trying to push the debate forward in a positive way. Nevertheless, the proposal itself raises several key worries. These at least require attention, and may ultimately undermine the whole approach (Gardiner 2017; 2021a; Arrhenius 2022). One worry is that, in context, encouraging the policy of "making the grandchildren pay" poses a profound threat of intergenerational extortion, or at least something very much like extortion.[3] However, there are also more specific worries that do not depend on that complaint. In this paper, I explore some of these latter concerns, in order to give them greater clarity and depth. For ease of presentation, I concentrate on Broome and Foley's World Climate Bank. However, many of the points being made apply much more widely, to proposals of the same general form. Hence, my account is of broad relevance to intergenerational ethics, whether concerning climate change or other issues.

One central conclusion is that we should be skeptical about optimistic arguments based on the mere possibility of Pareto improvements across generations. Among other things, mere possibility is not the right standard, Pareto-violating outcomes are plausible, and many Pareto-improvements are nevertheless ethically unattractive or even disturbing. More generally, "making the grandchildren pay" may encourage a tyranny of the contemporary and become mired in moral corruption. Indeed, at the extremes, it may overreach in ways which undermine norms of intergenerational ethics and threaten not only future catastrophe but also the gains

[3] Part of that argument is grounded on the claim that making the grandchildren pay is structurally similar to proposals mentioned or implied in other climate contexts, that those most vulnerable to severe climate impacts, such as poor countries (like Bangladesh and Haiti), should pay off the richer countries (such as the US, EU, China and Russia) to cease their excessive emissions (e.g., Posner and Weisbach 2010). Whereas making the most vulnerable pay suggests international climate extortion, making the future pay encourages intergenerational climate extortion.

of civilization and the existence of humanity itself. These are severe threats that ought not to be ignored by any serious effort at protecting the future. To highlight them, I contrast Broome's proposal for a World Climate Bank that implements "making the grandchildren pay" with my rival proposal for a Climate Justice World Bank that respects principles of global and intergenerational ethics.

1. Context

Earlier generations are typically in positions of asymmetric power over later generations, in ways that are easy to exploit for their own ends. Elsewhere I have explored this idea extensively, including by identifying a specific kind of intergenerational collective action problem that I call the tyranny of the contemporary (Gardiner 2011). In my view, the tyranny of the contemporary poses a basic standing threat to human communities and endeavors, and so should be at the heart of intergenerational ethics and political philosophy. I have also tried to show that the threat is live and severe in the case of climate change.[4]

One way to illustrate the tyranny of the contemporary is to point out that any given generation is in a position to make decisions about both front-loaded goods and back-loaded goods. For our purposes, let us say that front-loaded goods are those whose benefits accrue to the decision-making generation but whose burdens (e.g., costs, harms) fall on later generations; by contrast, back-loaded goods are those whose burdens come to the decision-making generation but which benefit later generations.[5] Notably, if the choices of a decision-making generation are driven primarily by its own generation-relative ends (or more narrowly by its own self-interest), we might expect it to oversupply front-loaded goods and undersupply back-loaded goods relative to wider ethical norms, such as those of justice, beneficence and virtue (Gardiner 2011).

To take a pure case, imagine that the current generation of decision-makers

[4] The idea is introduced and applied to climate change in Gardiner 2001, 2003, 2004, 2006, and elaborated in Gardiner 2011. A road towards a solution is proposed in Gardiner 2014, 2019, 2022b.

[5] The basic terminology of 'front-loaded goods' and 'backloaded goods' is intended to be generic, and to cover a variety of scenarios. Consider three especially salient examples. In one kind of case (call these, 'narrow cases') the decision-making generation accrues *only* benefits, and the later generations *only* burdens. In another kind of case (call these, 'wide cases'), the decision-making generation receives *net* benefits, and the later generations *net* burdens. In a third kind of case (call these, 'focused cases'), our interest is in benefits and burdens of specific types. So, for example, we might be concerned with situations where the decision-making generation receives morally trivial benefits (e.g., minor consumer luxuries), and the later generations morally serious burdens (e.g., major violations of basic rights). For illustrative purposes, narrow cases and focused cases are perhaps the most helpful, since they often present especially clear examples of intergenerational buck-passing (though wide cases can also be very compelling). In practice, when it comes to analyzing real world problems, wide and focused cases are likely to be the most prevalent.

can make a choice that will benefit them directly, but only modestly; however, the choice will inevitably bring on severe impacts 200 years in the future (e.g., negative climate impacts). Suppose that the choice would clearly be ruled out on ethical grounds (e.g., as profoundly unjust, maleficent, or callous), and that there are no unusual or confounding factors (i.e., the case is at it sounds). Still, the current generation of decision-makers may be tempted to make the ethically illegitimate choice. For example, suppose that it benefits them, has no direct costs for them, and only relatively minor indirect costs. Given this, to use terms I employ elsewhere, the decision-making generation is in a position to "pass the buck" for its behavior onto the future due to its asymmetrical power; it can engage in a distinctive kind of collective action problem, the tyranny of the contemporary (Gardiner 2011).

Unfortunately, it is all too easy for the threat of intergenerational buck-passing to be obscured. For the current generation, and especially the most affluent, are vulnerable to moral corruption: roughly-speaking, to ways of thinking and talking about problems which are seriously distorted and self-serving. Such framings facilitate continued exploitation of the future (Gardiner 2006, 2011, 2022a).[6]

Notably, there are various ways in which moral corruption can be encouraged and facilitated. These include complacency, selective attention, unreasonable doubt, pandering, false witness and hypocrisy. Perhaps the most obvious enabler of moral corruption is unreasonable doubt, which has been and continues to be prominent in the climate case. However, in this paper, I will concentrate on pandering. To pander is to "gratify or indulge (an immoral or distasteful desire, need, or habit or a person with such a desire)" (Lexico 2022). The proposal to create a World Climate Bank, with the specific objective of sparing the current generation from any sacrifice and making the grandchildren pay for climate mitigation, seems especially likely to engage this mode of moral corruption.

One reason I am interested in the proposal for a World Climate Bank is that Broome and I agree both that the current climate situation is very serious, and that new, intergenerational institutions are needed to resolve it effectively. However, beyond that we appear to have a sharp disagreement.

In my view, the right way to begin is by calling for a global constitutional convention on future generations (Gardiner 2014, 2019, 2022b). This would be a delibera-

[6] Those tempted to engage in bad intergenerational behavior are likely to favor ways of conceptualizing the situation that obscure what is really going on. This encourages distorted framings to emerge, particularly those which ignore or marginalize the intergenerational dimension, and cast the behavior of the current generation in a more favorable light (e.g., elsewhere I argue that the conventional prisoner's dilemma and tragedy of the commons analyses are helpful distortions from the point of view of moral corruption). Uptake of such framings is made easier by the lack of pressure on them coming from the victims, who, being young or not-yet-born, are poorly placed to question them, or to propose alternatives.

tive forum tasked with developing proposals for new, intergenerational institutions to confront the deep challenges humanity faces, now and over the very long term. Among other things, these new institutions would consider how to fund projects of intergenerational importance, such as climate mitigation. Importantly, they would do so in light of central intergenerational norms, including ethical norms.

By contrast, Broome believes that ethics has and will continue to fail us. Instead, he argues that climate policy should be grounded in direct appeals to self-interest, even if that results in injustice. He also appears to favor embedding his World Climate Bank in a fairly conventional institutional and political setting, since he emphasizes it on the existing World Bank and International Monetary Fund.

Part of the point of this paper is to explore these differences between us, and to clarify why I propose a more ambitious and thorough-going approach to institutional reform in the intergenerational context. As we shall see, a key issue is that Broome, like many writers on climate policy, appears not to regard the prospect of a tyranny of the contemporary as a real threat, or at least one worth emphasizing. I shall argue that this attitude manifests a troubling complacency. It also raises awkward questions about the possibility of moral corruption when it comes to how proposals like "making the grandchildren pay" are received and implemented.

2. The Standard Argument for "Making the Grandchildren Pay"

Initially, the main positive rationale offered for "making the grandchildren pay" ('MGP') appears straightforward. Relative to business as usual, advocates say, both the current generation of decision-makers and future generations can benefit from passing the costs of climate mitigation to the future: the current generation benefits since it no longer needs to absorb the costs of decarbonization; future generations benefit because they avoid severe climate change. Thus, "making the grandchildren pay" can deliver a strong Pareto improvement: both the current generation and future generations are made better off than under the status quo (and no one is made worse off). As Broome characterizes it, making the future pay generates a "win-win" scenario.

2.1. A Transformative Analysis?

In light of the Pareto argument, proponents of MGP typically assert that their argument is transformative: it radically alters how we should understand the climate problem, and thereby potential solutions. A first major claim involves the shape of the problem. Conventional accounts of how to allocate climate mitigation responsi-

bilities emphasize sacrifice, burden-sharing and requirements of justice[7]; by contrast, we are told, the Pareto framing implies that we should think of climate mitigation in terms of "the division of a surplus", and one that can be shared by the current generation and the future (Broome and Foley 2016).

A second major claim is that "making the grandchildren pay" can deliver efficiency without sacrifice. Under the Pareto argument, no generation need make a sacrifice, since each generation can be made better off than under the status quo. Notably, Broome himself appears to embrace two particularly strong forms of the "no sacrifice" position. In general, he says that "literally everyone" can benefit from pursuing climate action in a Pareto efficient way. Thus, it seems that there is no need to think in terms of burdens or burden sharing at all: "no one needs to bear a burden to mitigate climate change" (Broome and Foley 2022).[8] More specifically, Broome emphasizes the ability of the current generation to demand full compensation for itself for any threatened sacrifice. For instance, he says: "If we [the current generation] make a sacrifice by emitting less greenhouse gas, we can fully compensate ourselves by using more ... artificial and natural resources for ourselves. We can consume more, and invest less for the future" (Broome 2012, 44).

A third major claim typically made by proponents of MGP is that the strategy of "making the grandchildren pay" will enable us to move beyond the ongoing political inertia impeding climate action. Under "efficiency without sacrifice", advocates say, those in the current generation who would otherwise resist change can be placated. Future generations can "buy off" the recalcitrant (Broome and Foley 2016). They can shoulder the costs of climate mitigation themselves, in order to ensure that the current generation bears no burdens.

A fourth major claim is that "making the grandchildren pay" helpfully enables us to embrace self-interest rather than ethics. Broome, for instance, claims that emphasizing moral arguments has so far failed us in motivating climate action. By contrast, the "making the grandchildren pay" strategy allows us to move beyond ethical motivations to rely on the power of self-interest. As Broome puts it in his aptly-titled paper "Do Not Ask For Morality": "We should give up trying to solve the problem of climate change by appealing to the morality of governments. Instead we should concentrate on building the institutions that will make it possible to solve the problem without asking for morality" (Broome 2017).

---

[7] Conventional accounts likely include my own, together with those of Agarwal and Narain, Simon Caney, Axel Gosseries, Dale Jamieson, Catriona McKinnon, Henry Shue, Peter Singer, Olefemi Taiwo, and numerous others.

[8] Arrhenius points out that this is a high and implausible bar. He also points out that Broome and Foley's revised account of Pareto efficiency does not protect individuals in the future against suffering "no sacrifice" since it considers only collective resources and not population size. (See Arrhenius 2022; on the first point, see also Gardiner 2013).

A fifth major claim is that our practical priority should now become to develop institutions that will facilitate "efficiency without sacrifice". Enthusiasts argue that implementation of "making the grandchildren pay" will require institutional change, and that the institutions that are relevant here are primarily economic institutions (rather than, say, political or legal institutions).

The most prominent proposal is Broome and Foley's call for the creation of a genuinely new institution, the World Climate Bank, modelled along the lines of the World Bank and International Monetary Fund (Broome and Foley 2016). The World Climate Bank would be empowered to issue long-term debt on extended timeframes, of 300 years or so.

Other proposals involve a more modest reorientation of existing institutions. Thus, for example, Matthew Rendall argues that specific countries who are currently well-positioned in terms of debt, such as the Nordic countries and South Korea, should perform the more limited task of providing funds for the research and development of green technologies by taking on intergenerational debt (Rendall 2021).

A sixth major claim made by proponents of "making the grandchildren pay" is that the task of designing new institutions is one for economists. For instance, Broome says:

> Efficiency without sacrifice is technically possible. It is a big task for economics to make it practically possible. Economic institutions need to be created that can shift resources in the required direction. The economics profession should take on this responsibility. Making efficiency without sacrifice available would lubricate the political process, and make it much more likely that the problem of climate change will be resolved." (Broome 2012, 48)

Thus, Broome explicitly delegates the project of devising and developing his World Climate Bank to economics as a profession. Notably, he does not call on experts in ethics and justice, such as (say) those engaged in designing human rights regimes, constitutions, development, or humanitarian aid. Whereas economists are presumed to have the right expertise, those in (say) philosophy, political science, law, development, human rights organizations, and so on, do not even rate a mention.

A seventh major claim suggests one reason for this. Typically, advocates for "making the grandchildren pay" maintain that conventional concerns about justice should be largely set aside.

At a theoretical level, we see two main camps (cf. Gardiner 2017). On the one hand, some proponents of MGP (e.g., Broome 2017) adopt a concessive approach: they grant that their approach is morally problematic, in part because it perpetuates injustice and violates duties of beneficence. Nonetheless, they argue, "the best"

should not become "the enemy of the good": the stakes are high enough that we should pursue MGP anyway. On the other hand, other proponents of MGP (e.g., Rendall 2021) appear to adopt a more enthusiastic theoretical understanding of "making the grandchildren pay": they suggest that the relations between generations are such that MGP raises little or no moral problem. Both camps emphasize the expectation (coming from mainstream economics, they say) that the future will be much richer, and so better off, than the present.

At a practical level, one symptom of the marginalization of justice is that both the concessive and enthusiastic camps approach institutional proposals with notable enthusiasm and few, if any, reservations. For instance, Rendall says:

> The moral case for EWS [efficiency without sacrifice] is strong. If states can agree to establish a world climate bank, as Broome and Foley propose, they should do it.... States need have no reservations about financing a research programme to discover such technology through public debt, in part or in full. It will be better for future generations to share in the effort than for us to do too little, too late." (Rendall 2021, 980)

Similarly, Broome and Foley urge:

> Provided the quantities are properly balanced, future generations will end up better off: the cleaner atmosphere will more than compensate for the smaller quantity of conventional capital they receive. We know that a proper balance can be found, because the theory tells us that a Pareto improvement is possible. (Broome and Foley 2016, 163)

> Is it institutionally feasible to issue all this debt? It certainly should be. We have shown how a Pareto improvement can be achieved in the real economy. Conventional investment needs to be shifted into reducing greenhouse gas emissions. The only remaining question is whether the world's financial system can make it happen. It would be a terrible indictment on the world order if the great gains that could be achieved by controlling climate change were prevented by the weakness of the financial system. (Broome and Foley 2016, 166–167)

In summary, the standard argument for "making the grandchildren pay" paints a comparatively rosy picture of our predicament and how to find a feasible way out. The central ideas are: we should see the climate problem as one of sharing a surplus, and facilitating Pareto improvements; under these conditions, everyone benefits and no one needs to make any sacrifice; this enables us to overcome ongoing political

inertia by embracing the power of self-interest; the main difficulty is practical and concerns implementation; but the solution is relatively straightforward: economists should design new, but more-or-less conventional institutions that allow the costs of climate action to be passed on to future generations.

This rosy picture stands in stark opposition to more conventional accounts of both the climate problem and the needed solutions. For instance, conventional diagnoses typically emphasize deep issues of injustice, institutional inadequacy and moral failure; similarly, conventional solutions often require sacrifice, major structural reform, and wrestling with wider injustice on a global scale. By contrast, "making the grandchildren pay" offers the promise of setting aside such issues. Initially at least, it appears to reveal a simpler and more political tractable strategy for making progress. Consequently, it is bound to sound attractive to some.

2.2. Overview of My Response

Alas, I fear that this rosy picture is misleading, and in ways which obscure much of what is really at stake. We might even say that it "sugarcoats" where we are (Gardiner 2017). In the subsequent sections, I will summarize some central concerns and extend them. But before diving in, let me offer a brief overview of my response.

First, I am critical of the framing. In particular, the standard argument fails to acknowledge (i) the serious threats to the future posed by climate policies based on "making the grandchildren pay", (ii) the grim reality of many potential Pareto "improvements", and (iii) the wider risks of transforming global norms in ways that threaten both future generations and humanity's progress more generally.

Second, I am concerned about the policy proposals. Importantly, the framing problems complicate, and may ultimately compromise, the institutional proposal for a World Climate Bank. In particular, (iv) to guard against the obvious threats, such an institution would need to embrace strong intergenerational ethical norms; (v) these norms go far beyond the call for Pareto improvements: and (vi) are themselves in need of further elaboration and defense. Moreover, (vii) a climate world bank, together with the global institutional architecture in which it resides, would also have to be capable of delivering on such norms, including not just in its own decisions but also in wider enforcement of those decisions. None of this seems easy, and (viii) much of it invites the re-emergence of many of the obstacles to effective climate action that have hampered earlier international efforts.

Third, these concerns have implications for how to proceed. One consequence is (ix) that it is far from obvious that Broome's own World Climate Bank is fit for purpose, especially since it is to be structured along the lines of current institutions such as the World Bank and IMF, and supported by national governments as con-

ventionally understood. Another is (x) that, arguably, making it so requires more ambitious proposals for institutional reform, such as my proposal for a global constitutional convention for future generations (Gardiner 2014, 2019). Indeed, (xi) without such reform, Broome's World Climate Bank and similar proposals may well turn out to be more a further manifestation of threats to future generations, rather than an effective foil to such threats.

In section 4, I shall develop these arguments through four basic points. However, before doing so, let us clarify the terms of the debate.

3. Terminology

3.1. "Making the Grandchildren Pay"

The first clarification concerns the slogan "making the grandchildren pay". This is primarily a catchphrase: a short expression that indicates a wider idea in a memorable way. Importantly, the invocation of 'the grandchildren' is intended to be eye-catching rather than literal. For one thing, the proposal is not restricted to grandchildren as such (whatever that might be supposed to mean).[9] Instead, the focus is on future generations broadly conceived, where this includes young people alive today and those not yet born. In essence, the idea being signaled is that young people and those not yet born (or some subset of them) should pay for climate mitigation, as a way of "buying off" or "compensating" the current generation of decision-makers in exchange for their cooperation in climate action. Thus, the phrase 'the grandchildren' stands in for (loosely-speaking) all those future generations who would be benefited by mitigation in the longer term, and so positively impacted by making the bargain with the current generation of decision-makers. Plausibly, this means most of those who live beyond the lifetimes of current decision-makers, stretching at least a couple of hundreds of years into the future, and perhaps much farther.[10] Given this,

[9] Numerous issues would arise if it were so restricted. Consider just two examples. First, one proposal would be that the burdens should be borne only by those whose grandparents are currently living, and so count as "grandchildren" by that definition. Second, another proposal would be that members of the older generation could be bought off only if they themselves were grandparents, rather than (say) childless. Such proposals might be of interest if "making the grandchildren pay" was understood in a restrictive way. However, since that is clearly not the intention of authors like Broome, I set it aside here.

[10] None of this has much to do with "grandchildren" as such. For one thing, many of these are not "grandchildren" of those currently alive – some are children, some will be great grandchildren, some great-great-grandchildren, and so on. More generally, the crucial point has nothing to do with family lines. Some of those alive now will have no descendants, and so none of those being asked to bear burdens in the future are "grandchildren" of theirs, or indeed any kind of direct descendant. More importantly, one's own immediate familial connections with the future will only constitute a very small percentage of the vast number of people who will actually be affected by choices about climate change

a more accurate label than "making the grandchildren pay" would be something like "making the otherwise climate-burdened future generations pay", or in abbreviated form, "making the future pay" (MFP).[11] So, in what follows I shall use this phrase instead.

3.2. Contribution vs. Absorption

The second clarification concerns what is meant by 'pay' in the phrase "making the future pay". As it stands, this is ambiguous and in ways that are likely to cause confusion.

First, taken out of context, the phrase "making the future pay" might be understood in a very weak or minimal way, as meaning only that the future should contribute something to the project of climate mitigation, as part of some intergenerational burden-sharing scheme. We might call this, 'the contribution principle', and take it to refer to the general idea of "making the future contribute (something)" to the costs of climate action ('MFC'). I will consider the idea of MFC later in the paper. For now, let us simply observe that the contribution principle is clearly not the principle in operation in our context.

Instead, in arguments like Broome's the internal logic of "making the grandchildren pay" is that the future must absorb all of the costs of climate mitigation, or at least all of those that would otherwise be passed on to the current generation (see later). The central point of Broome's proposal is that the current generation must be spared any sacrifice. Importantly, his version of MGP is not advocating a burden-sharing scheme to which each generation must make a contribution. On the contrary, its core goal is to shield the current generation, so that they make no sacrifice. Thus, this second version of "making the future pay" endorses the 'absorption principle': the key idea is to spare the present by making the future absorb the costs of climate mitigation.

It is also worth mentioning a third interpretation of the phrase 'making the future pay'. In English, one popular idiom involves someone using the form of words "I will make X pay" as a way of signaling that a state of enmity exists between themselves and X, together with their intention to punish X. The key idea in play then is that of retribution against X, typically (though not always) for something that X has done. We might call this 'the retributive principle'.

Clearly, this is not the sense of "making X pay" operative in our context. Propo-

---

and shifting burdens. Thus, the "grandchildren" framing tends to obscure the scope of the implications for unethical action toward the future.

[11] I shall ignore many of the complexities that arise in this setting, such as issues of generational overlap, nonexistence, and the nonidentity problem.

nents of "making the grandchildren pay" have no such enmity, and in no way intend to punish future generations. Nevertheless, it is perhaps worth noting that critics of MFP tend to worry that some versions of it turn out to be both unduly harsh on future generations, and to undermine the right kind of intergenerational relationships. While not signaling that the current generation and the future are enemies, some forms of "making the grandchildren pay" appear to express acquisitive, adversarial, or even manipulative attitudes towards the future, and these threaten to undermine or destroy any bonds of fellowship, solidarity or even minimal good will between the relevant generations. More generally, there is a background worry that some proposals that rest on the absorption principle are liable to alienate generations from one another, perhaps in a particularly deep way. One example of potential alienation is when "making the grandchildren pay" appears extortionate (e.g., Gardiner 2017); but there are others.

3.3. 'World Climate Bank'

Our third terminological clarification concerns the phrase 'World Climate Bank'. This should be approached with care. The mere form of words 'world climate bank' communicates only: (a) a single body; (b) global in scope; (c) focused on climate alone; which (d) takes the particular institutional form of a bank. Nevertheless, in the context of Broome's argument, 'world climate bank' clearly means something much more specific, necessitating further claims. In addition to features (a)-(d), Broome's 'World Climate Bank' also: (e) has the central aim of foisting the costs of mitigating climate change onto future generations; (f) is charged with achieving this aim in such a way that ensures that the current generation experiences no sacrifice; and (g) is structured so as to relevantly similar in nature (e.g., in character and design) to the International Monetary Fund and existing World Bank. Given that this specificity is obscured by the much more generic label 'world climate bank', it seems wise to introduce some further terminology.

Climate-specific financial mechanisms

Let us begin with the bare idea of a financial mechanism focused on climate. Call this a 'climate-specific financial mechanism' ('CFM'). Such mechanisms might function at various levels, including local, national, regional, and global levels. So, for instance, we might distinguish 'national climate-specific financial mechanisms' (NCFMs) and 'global climate-specific financial mechanisms' (GCFMs). Given this, we can already see ways in which Broome's World Climate Bank is controversial. Consider three examples.

First, the general proposal to pursue climate-specific financial mechanisms is

not obviously correct. Some may favor financial institutions that are either broader or narrower in their scope.[12] So, for example, on the broader side they may advocate for more general environmentally-focused institutions rather than only climate-specific ones (e.g., an Environmental World Bank); on the narrower side, they may propose financial institutions that are linked to some aspects of climate policy (e.g., to technological transfers, adaptation or loss and damage), rather than others (e.g., a Green Technology World Bank).

The second area of controversy involves the gap between the appeal of having climate-specific financial institutions of some sort and Broome's specific proposal. Though many might favor some form of CFM, they may do so for a wide variety of reasons. Crucially, simply advocating for climate-specific financial mechanisms in no way puts one in the position of endorsing "making the future pay", or "efficiency without sacrifice", or specific institutions designed to implement these principles, such as Broome's World Climate Bank.

This point is worth emphasizing. While proponents of MFP do advocate for climate-specific financial mechanisms, theirs are CFMs of a highly-specific, and indeed controversial, form. The specific form rests on the distribution of burdens across people over time, and especially across generations. MFP demands (i) that the future shoulder all the burdens of a climate transition, and (ii) that the current generation suffer no sacrifice. We might say that MFP "saddles the future", and "spares the present". Hence, the climate-specific financial mechanisms advocated for by advocates of "making the grandchildren pay" are restricted to those that are essentially both future-saddling and present-sparing.

Notably, this restricted vision of climate-specific financial mechanisms contrasts with other possible approaches to organizing CFMs. As we have seen, one obvious approach is one where burdens are shared over time, so that different temporal groups adopt a contribution principle based on intergenerational burden-sharing. Another approach would be one where the current generation absorbs the burden and spares the future. Either alternative approach might turn out to be ethically justifiable, and that would need to be the subject of further argument and analysis. Presumably, some versions of each would also be ethically unjustifiable or ethically flawed; and that would be worth investigating too.

These are matters from another time. The point here is simply that there are genuine rivals to "making the future pay" when it comes to thinking about how to design climate-specific financial mechanisms. Thus, it is important that proposals

[12] Of course, some will be against any kind of climate action or genuinely global institution; and others will say that genuinely global institutions are needed if we are to have a good chance of resolving the climate crisis in any ethically reasonable way. But such matters are not my concern in this paper (see Gardiner 2014, 2019).

based on MFP, like Broome's "World Climate Bank", do not get a free pass simply because they advocate for some form of climate-specific financial mechanism. Many other approaches might do the same, including those that argue for burden-sharing rather than saddling the future. Crucially, Broome's proposal is not neutral in this regard: that it rejects such proposals – and as part of its basic rationale – is a central feature of his World Climate Bank.

The third area of controversy concerns the appropriate institutional models. Broome's proposal is for a new institution designed by economists and modelled on the existing International Monetary Fund and World Bank. However, one difficulty here is that the IMF and WB have serious critics. Some object that they are strongly influenced by a small number of very powerful actors, and so (among other things) run counter to acceptable norms of justice and legitimacy. Other critics are against any kind of genuinely global institution, and more sympathetic to financial institutions grounded at a national or regional level.[13]

Another kind of difficulty is the seemingly exclusive reliance on economists. Many would argue that other kinds of expertise would be relevant to designing a successful climate-specific financial institution, such as in ethics, law, governance, and so on.[14] Of particular concern might be expertise relevant to forming institutions that are appropriately inspired by, grounded in, reflective of, and accountable to, various principles and ideals of justice.

Climate Justice World Banks

This leads to a second terminological point about climate world banks. Given that we are concerned with ethics and justice, let us call climate-specific financial mechanisms (CFMs) that seek to allocate burdens according to reasonable ethical guidelines, principles or requirements, Ethical Climate-specific Financial Mechanisms (Ethical CFMs), and those that take the form of banks, Ethical Climate Banking Mechanisms (or Ethical CBMs). A subset of Ethical CBMs will be those that satisfy requirements of justice. So, we might refer to these as Just CBMs. To fix ideas, and give us a simpler label to work with, let us focus on Just Climate Banking Mechanisms at the global level, and call these Climate Justice World Banks.

Now, our central concern is with intergenerational issues. Hence, for current

[13] Notably, Rendall (2021), for example, while advocating for MFP, seems sometimes to favor national actors rather than a world climate bank.

[14] In a related comment, Broome and Foley say that "the *managers* of a WCB would have the responsibility of detecting and rejecting fraudulent or misleading applications for loans based on expenditures that in fact would have no impact on greenhouse gas emissions" (Broome and Foley 2022, 11; emphasis added). This seems surprisingly weak kind of enforcement mechanism. I would have thought that international law, powerful courts, anti-corruption organizations and a relevant police force should be considered as well.

purposes our focus will be on Intergenerational Climate Banking Mechanisms ('ICBMs') or (for short) 'Intergenerational Climate Banks'. Some of these will be in accordance with justice, and so can be called Just ICBMs. Just ICBMs will presumptively fall under the general category of Climate Justice World Banks. Some may be embedded within a wider institution that we might call a Global Justice World Bank.

Notably, the relationship between "making the future absorb" the costs of climate mitigation and the idea of a Climate Justice World Bank is not simple or obvious, as we have already seen. For instance, on the one hand, some proponents of "making the future pay", such as Broome, explicitly concede that their future-saddling ICBM would perpetuate injustice. So, they do not think of their "World Climate Bank" as one form of Climate Justice World Bank. However, on the other hand, some proponents of MFP maintain that saddling future generations with the burdens of climate change would be in accordance with justice (or could be under some circumstances). Hence, they are interested in the idea that MFP might be endorsed by some form of Climate Justice World Bank.[15]

Summing up, in this section we discussed ways in which the language employed in the debate about the "World Climate Bank" is unhelpful and potentially misleading. To help clarify some crucial issues, we introduced some new terminology. Most notably, this included: 'making the future pay' (as opposed to 'making the grandchildren pay'); 'making the future absorb' and 'making the future contribute' (as opposed to 'making the future pay'); 'climate-specific financial institutions' (as opposed to 'World Climate Bank'); 'Intergenerational Climate Bank'; 'Climate Justice World Bank'; and 'Global Justice World Bank'.

The hope is that distinguishing these different concepts helps to promote a broader vision of the possibilities, and to highlight what is most distinctive and controversial about Broome's own proposal.

4. Four Basic Points

With these ideas as background, let us now assess the general idea of "making the future absorb" the costs of climate mitigation thorough the creation of an Intergenerational Climate Bank, and especially one grounded in "efficiency without sacrifice", self-interest, and perpetuating injustice. I offer four basic points.

[15] Relatedly, Broome himself does say that his future-saddling World Climate Bank at least has a moral purpose. Hence, he may believe that there are some limits on how *unethical* a MFP policy one should adopt. Hence, for example, he might think that his WCB must be at least tolerably unjust, and not fall into complete moral horror.

4.1. The Playing Field of Possibility

The first basic point is that much of the argument for MFP is conducted on what I shall call "the playing field of possibility". Proponents treat establishing the possibility of Pareto improvements as their key move, and then move quickly from that to institutional proposals. For instance, Broome and Foley say:

> The current generation leaves greenhouse gas for future generations, but it also leaves them nice things. It leaves conventional capital such as roads and cities, and it leaves natural resources, because it does not use up all the natural resources it could. It can therefore compensate itself for reducing its emissions of greenhouse gas. By reducing its transfer of resources forward in time, it can in effect transfer resources backwards from future generations to itself. This transfer can serve as compensation from future generations to the present. In effect, the current generation has only to switch some of its investment from building conventional capital to reducing greenhouse-gas emissions. By this means, a Pareto improvement is possible. (Broome and Foley 2016, 158)

> Future generations benefit from the current generation's conventional investment, and there will be less of that, but they will gain a cleaner atmosphere instead. Provided the quantities are properly balanced, future generations will end up better off: the cleaner atmosphere will more than compensate for the smaller quantity of conventional capital they receive. We know that a proper balance can be found, because the theory tells us that a Pareto improvement is possible (Broome and Foley 2016, 163)

> The theory tells us that a Pareto improvement is possible. How, in more detail, can it be achieved? (Broome and Foley 2016, 163)

As these quotes make clear, the words "can" and "possible" are at the heart of how the central Pareto arguments are presented.

Unfortunately, this style of argument raises concerns. Most generally, establishing bare possibility seems a very low bar, epistemically and practically. Epistemically, it is typically fairly easy for philosophers and economists to agree that something is possible. However, this shows little. For one thing, it is very permissive: many things are possible. For another thing, philosophers and economists are often accused as having unduly relaxed ideas of what is possible, especially by their colleagues in the policy world; this is particularly so when the arguments for possibility are made on theoretical grounds.

From the point of view of action, mere possibility is also a low bar. We do not (and

should not) organize policy around pursuing outcomes merely because they are possible. The principle "if a good outcome is possible, we should develop institutions to facilitate that outcome" is not a compelling principle. Again, many good things are possible, and we cannot pursue them all. Moreover, policies that make good things possible often also make bad things likely. So, a more general assessment is needed.

In short, establishing bare possibility does not get us very far. Although many may be willing to admit that attractive MFP policies (including a desirable form of intergenerational climate bank) are possible in theory, this falls well short of what needs to be shown to justify an actual policy initiative.

Instead, what is needed is to show that the MFP approach is especially salient. The outcomes it pursues must be realistic, and likely (perhaps very likely) to be achieved through building the relevant institutions and implementing the related policies. Moreover, the MFP approach must look attractive relative to alternatives. If in fact MFP actively encourages negative outcomes that are also salient, or if other strategies are more likely to work, then claims of theoretical possibility will quickly lose their interest.

Such criteria are relevant in the climate context. One reason is that there are other salient possibilities, and these may undermine the appeal of MFP and point in a different direction (see below). Another reason is that the key feature of MFP that is emphasized by its proponents – that Pareto improvements are possible – is not the only standard against which options can be judged, and may not be the most important. For example, additional standards include those of encouraging justice, protecting welfare over the long-term, and not encouraging or facilitating extortion.

Of course, proponents of MFP believe that their proposals are realistic, and indeed this appeal to realism is central to their case. Specifically, they maintain that MFP has the major political advantage of appealing to self-interest of the current generation as a way of unlocking political inertia. However, to make good on such a claim we would need further arguments, beyond the mere possibility of Pareto improvements. For, as we shall now see, there are other salient possibilities.

4.2. Pareto-Violating Cases

My second basic point is that one salient category is that of Pareto-violating cases: situations where at least some are made worse off, perhaps substantially worse off. Such outcomes are surely possible. They also seem clearly salient in the climate case, and indeed more generally when it comes to relations with the future. For instance, advocates for future generations will be concerned about cases where the current generation runs up intergenerational debt in the name of MFP, but ultimately does not deliver the expected benefits to the future.

One obvious example would be if the current generation sets up a World Climate Bank, draws loans from it under the pretext of enhancing climate mitigation, but then ultimately diverts those funds to other purposes, such as (say) for short-term consumption for its own benefit. Abuses of funding programs are, after all, hardly unprecedented in other areas.

Less obvious, but still plausible, examples would include cases where the funds are used for projects that would have proceeded anyway, or to fund low-hanging fruit with significant co-benefits for the current generation, and where this may be done in ways that are not sustainable in the long-term. Those familiar with the difficulties that have confronted other climate policies will notice that there is ample precedent for such shenanigans (e.g., the CDM; REDD; offsetting).

Of course, the standard argument assumes that the policies that will be pursued under MFP are only those that involve genuine Pareto improvements. Yet this is merely an assumption, nothing more. Crucially, there is no magic involved in pursuing policies of MFP that guarantees that doing so will actually serve to make the future better off. As the other examples mentioned above (CDM, REDD, offsetting) show, having high hopes for a climate mechanism is not enough to ensure that it will deliver. Among other things, generational bad faith is possible (and perhaps likely, given the problems of intergenerational buck-passing and moral corruption). Moreover, even if at the outset there is generational good faith, so that early proponents of MFP genuinely want to benefit the future (as I am confident that current philosophers and economists who advocate MFP and a Broomean World Climate Bank do), what those who initially propose policies intend and what actually results (in practice and over the longer term) are two different things. For example, good policy ideas are often hijacked by powerful interests; sometimes this happens quickly; sometimes it occurs slowly and inexorably over time.

In summary, in my view, the possibility of Pareto-violating outcomes remains salient even when we notice (with the standard argument) that Pareto outcomes are possible. In particular, advocates for the future are right to be suspicious of institutions that do not take seriously the threat of outcomes that make future generations worse off, including worse off than under the status quo or "business-as-usual" baselines.[16] There is too much at stake to be complacent (and too much history of intergenerational buck-passing and moral corruption in international climate policy).

[16] Advocates for marginalized populations in the current generation should also be concerned about non-Pareto outcomes from them. There are risks of *intra*-generational buck-passing, and of moral corruption, here too. Still, the prospects of those risks being noticed and resisted are at least a little higher.

4.3. Disturbing Pareto Baselines

My third basic point concerns another category of salient possibilities: cases of MFP that technically count as Pareto improvements, but nonetheless remain deeply morally troubling.[17]

The Older, Decision-Making Generation

Consider first the relevant Pareto baseline (i.e., the baseline from the point of view of strictly generational self-interest of a simple, narrow sort) for the older, decision-making generation: say those aged 50–80. This baseline is likely to be high (and perhaps very high – see section 4.4).[18]

On the one hand, the benefits of keeping close to the status quo over the next few decades are likely to remain appealing to the older generation. In the short- to medium-term, the status quo largely preserves ways of life to which they are accustomed. Transition is risky, and likely to be costly. This may seem particularly true of the kind of radical, "unprecedented" transition mainstream scientific reports say is now needed for even moderate chances of achieving the 1.5°C and 2°C targets (e.g., IPCC 2018). This attitude may also be especially prominent among those with the most resources and political power in this generation, whose influence makes the biggest difference in practice.

On the other hand, when thinking of the downsides of the status quo – bad climate impacts – there are considerable time-lags which are likely to be highly salient to the older generation (again, solely from the point of view of generational self-interest). For one thing, even if they would prefer to avert the worst climate effects, for the older generation "avoiding the worst" is complicated by the fact that many of those effects are relevantly distant: they will not arise for at least decades, often centuries, and sometimes thousands of years into the future; in short, not until they are dead and gone. For another thing, given the time-lags, even if some bad impacts will occur relatively soon – over the next few decades, while they expect still to be alive – many of these are already "in the cards", in the sense that they cannot

[17] Of course, there are other, more general reasons to be troubled by those actors (as opposed to theorists) who take the Pareto baselines to be especially salient. For instance, arguably, even taking the point of view of generational self-interest to be paramount in the climate case itself *expresses a profound disrespect* for future generations (e.g., through indifference to their plight, or excessive concern for one's own generation, or an attachment to lesser values (such as conspicuous consumption), or for other reasons). But here my concerns are more specific to the standard argument.

[18] For current purposes, I assume we are talking about simple, narrow and primarily economic understandings of self-interest, since I presume that this is what Broome and others have in mind in their "pragmatic", nonmoral arguments. I myself would reject such a conception of self-interest. For discussion, see my analysis of the similar narrow arguments offered by David Weisbach (e.g., Gardiner 2021b).

be avoided, or at least avoided through emissions reductions that are likely to prove physically or socially feasible (Gardiner 2011, In press). For both reasons (relevant distance; "on the cards"), it seems likely that much climate mitigation does not benefit the current generation of decision-makers directly.

Given the pressure from each side (the appeal of the status quo; the influence of time-lags), if we assume that the older generation's main relevant motivation is generational self-interest (per the standard argument for MFP), it is likely to demand more "compensation" for deviations from business-as-usual than one might initially think, or than would younger generations. The baseline against which a strong Pareto strategy must benefit the current generation of decision-makers is likely to be high. (In section 4.4, we see that it may be very high.)

More generally, we must beware of the comfortable language of "efficiency without sacrifice" and Pareto improvement. On the surface, both phrases are compatible with – and indeed tend to highlight – the idea that the current generation of decision-makers is merely made no worse off than under the status quo. However, the wider shape of the MFP argument goes further: it emphasizes the claim that the current generation will actually benefit from MFP, and so undergo an improvement in its condition; moreover, this seems central to the idea that MFP appeals to self-interested motivation, rather than morality. In essence, then, the key motivating thought is not "efficiency without sacrifice" but efficiency with profit (Gardiner 2017). Without some positive motivation that gives the current generation reason not to prefer the status quo, the self-interested argument cannot get off the ground. (Perhaps an ethical argument could fill this gap – see below – but this would undermine one central appeal of MFP highlighted by its proponents.)

Even more importantly, if self-interest is the key motivation, then it seems that the relevant baseline should not be "no sacrifice" or what the current generation expects without MFP. Instead, it should be whatever they can realistically get away with demanding. However, unfortunately this baseline is likely to be very high. The simple reason is that it will be the current generation itself that first implements MFP and creates the new intergenerational climate banks. It therefore has profound asymmetric power over the situation. Thus, at first glance, it is not clear what would prevent it from aiming to maximize its possible gains from the venture. As we shall now see, that appears to be a very high baseline; it is also likely to be a morally horrifying baseline.

The Young and Other Future Generations

Consider now the Pareto baseline for future generations. This is likely to be low; indeed, dangerously low. For the salient threshold there is making the future gen-

erations better off than they would be in a genuine climate catastrophe; but this is, by definition, a very bad state of affairs.

Many of the multiple threats posed by climate change are clearly catastrophic on a range of scales (e.g., local, national, regional, global). They also become manifest at different levels of global warming, some of which are uncertain. At the higher end, plausible catastrophes include scenarios that involve rises in global average temperature of somewhere in the region of 4°C by the end of the century, where some say that this threatens the collapse of global agriculture, and so the food supply of humanity.[19] Mainstream science also suggests that we should at least consider threats of even more extreme warming, of (say) 5-6°C or more, which would potentially undermine conventional social systems altogether and render large parts of the world uninhabitable for human beings. These are just two of the salient possibilities. They suggest that many of the relevant Pareto baselines for future generations are very low. They may bring on truly dystopian circumstances, such as those depicted in Cormac McCarthy's *The Road* or Margaret Atwell's *MaddAddam trilogy*. Indeed, both authors were partially inspired by the climate threat.

This brings us to a crucial point: making future generations better off than under baselines of genuine global catastrophe is (i) not very demanding, and so (ii) barely a constraint at all. For one thing, it is not very demanding because satisfying the Pareto criterion for future generations can be met merely by bequeathing them a condition slightly better for them than genuine global catastrophe. Yet this is compatible with bequeathing all manner of severe, but still lesser, circumstances to the future.

Moreover, in context this threatens to render the Pareto principle barely a meaningful constraint at all. Consider just one salient scenario. Suppose the current generation were to hand on to the future a climate 3–4°C warmer and saddled with heavy intergenerational debt. Climate change in that world is severe, and well above the conventional thresholds of 1.5° and 2°C for dangerous anthropogenic interference. Suppose that the combination of climate damages and inherited debt were sufficient to depress the quality of life well below that common in affluent countries now. Imagine, for instance, that it would set back human progress by hundreds of years. Still, this might be "better than genuine climate catastrophe" for the future, and so satisfy the Pareto principle, and be justified by the standard argument for MFP and the Broome's World Climate Bank.

This result appears deeply objectionable. For one thing, intuitively, "making the

[19] By 2100, the IPCC suggests temperature rises of 3.3-5.7°C under a very high emission scenario and 2.1-3.5°C under an intermediate scenario (IPCC, 2021, 17); by 2300, the projections are 6.6–14.1°C under a very high emission scenario, and 2.3–4.6°C under an intermediate scenario (IPCC, 2021, 2–10).

grandchildren pay" in such ways seems seriously wrong. For instance, it seems to violate any reasonable duty of care we might have toward future generations, especially given that other options are on the table. For another thing, painting this and similar scenarios as a simple "win-win" is deeply misleading, and obscures what it at stake, ethically-speaking. For instance, consider a suggestion from the current generation of the form "we'll hand you "only" a 3.7°C rise – and so spare you complete agricultural collapse – in exchange for you accepting a huge intergenerational debt". Proposing such a bargain seems much more accurately cast as intergenerational extortion than as a "mutually beneficial transaction".

Pressure on the Pareto Argument

This last point has wider relevance (Gardiner 2017). On the one hand, future generations and their advocates may not accept extortionate bargains. Some people resist extortion. They do this even in cases where the extortionate bargains look to be Pareto improving, and so good in one respect. This is not obviously irrational. For one thing, welfare is not the only value, and can be outweighed by others; for another thing, acceding to extortion might be beneficial in the short-term, but undermine welfare in the longer-term. (These points are quick to state, but nevertheless of central importance.)

On the other hand, even if future generations would accept an extortionate Pareto improvement, they would be doing so under compromised circumstances. A realistic threat of catastrophic climate change is being made against them. Their choice is therefore made under duress. Consent under extortion is morally compromised, and has little or no moral value.

In short, in context, the claim that future generations would accept the bargain being offered by the current generation becomes seriously misleading. It may not be true; and if true, it is likely of limited moral relevance.

Summing up, once we notice the different baselines for the current generation (high) and for future generations (low), the achievement of Pareto superiority is cast in a different light. Rather than Pareto appearing evenhanded – "both generations benefit!" – the asymmetry in relevant baselines implies a deep bias toward the current generation and against the future. Crucially, Pareto alone offers little to no protection to future generations ("the grandchildren"). The baseline seems not only low, but dangerously low. As a result, the standard argument for MFP opens the future up to all kinds of treatment that otherwise appears objectionable. In my view, these are highly salient possibilities that any proposal for MFP, or concrete policy recommendations like establishing an intergenerational climate bank, must take seriously.

4.4. Overreach

This brings us to my fourth basic point, which concerns the persistence of perverse incentives. The greatest risk of all may be that "making the future pay" and the new institutions charged with implementing it themselves pose a dramatic new threat to the future in terms of their potential for overreach. Once norms are established that legitimize MFP, and institutions that can make it happen, then new rationales for more payments can easily emerge.

Most notably, what ensures that the new system based on MFP will not itself become an agent of, or otherwise facilitate, intergenerational buck-passing? The temptations are clearly still there. The current generation can use the new institutions to take more from the future than is justified, and also benefit the future less than is needed. Once more, it is not clear what would stop them. Again, those not yet born are unavailable to play a substantial role in running the institutions, or holding them accountable, while the young probably lack the necessary social power as things stand.

Overreach may occur within the climate domain. The initial official remit for MFP is financing mitigation, and especially research and deployment of energy technologies. But why would the current generation stop there? Many other "burdens" can be connected to climate. Broome, for example, has already suggested that MFP be used to compensate for historical climate emissions, even though these are not matters of climate mitigation (Broome 2017, 17). Presumably, it won't be long before we see further demands: for instance, that the future pay for adaptation measures, loss and damage, relocation of climate refugees, ecological restoration, and much else besides. Once launched, what is to stop Broome's future-saddling World Climate Bank from rapidly expanding beyond mitigation to cover all these other areas and more?

Yet climate overreach is only the tip of the iceberg. Pressure for further expansion seems likely to arise much more generally. In time, we might see demands that the future should pay to address all manner of other threats that arise to them, for example, for nuclear disarmament, pandemic preparation, toxic waste disposal, and so on. Once there is a World Climate Bank, how long before we see calls for "making the future absorb" the costs of many other things too: for example, through a World Disarmament Bank, a Global Biodiversity Bank, a Weapons of Mass Destruction Bank, A Global Reforestation Bank, and so on? Indeed, ultimately, there may be calls for a grand simplification. Plausibly, the current generation will end up arguing that they need an overarching Future-Saddling and Present-Sparing institution, such as a Broomean Intergenerational World Bank or Intergenerational National Bank, for them to call on whenever the "need" strikes them, for whatever purposes they choose.

Over time, it is not difficult to imagine a major erosion of intergenerational ethical norms, as we slide toward making the future pay for a huge range of things formerly thought to be the responsibility of each individual current generation in turn. Eventually financial institutions based on MFP may simply become unconstrained revolving lines of credit for whatever the current generation sees fit to demand funds to do. Of course, the mechanisms of moral corruption are such that such activities won't be framed in these ways. But that does not ensure that the framing is not apt.

These various possibilities are already morally shocking. They pose a profound threat to future generations. Nevertheless, there is a further possibility that is even worse. MFP writ large also puts at risk humanity's progress. If institutions are set in place that license MFP to such an extent that the present generation at any particular time can effectively claw back from the future whatever it likes, this puts at risk the accumulation of the gains of civilization. It also makes humanity in the future more vulnerable to existential risks. In my view, these are at least among the salient possibilities that should be considered in any serious assessment of the MFP approach and the institutions empowered to deliver it. They may even be the most salient possibilities.

Again, recognizing the perverse incentives invites another comparison to paradigm cases of extortion. People say that once an extortionist has his claws into you, he never stops, but keeps coming back for more. This is a central reason why some resist extortion even when acceding to the initial threat appears to represent a strong Pareto improvement for them. For to accede may be merely the first step in a downward spiral, with no end in sight. Here, the initial Pareto improvement is at best short-lived, and in the end short-sighted. Given this phenomenon, one can see why defenses of MFP and a Broomean WCB might be resisted and for good reason. Claims about the theoretical possibility of strong Pareto improvements, or "win-win", fail to take seriously the underlying risks of the situation. (As Tony Soprano says, "I won't pay; I know too much about extortion" (cf. Gardiner 2017).)

One might even go so far as to argue that worries about intergenerational extortion and the threat to humanity's progress make resisting mechanisms of MFP itself a high priority for intergenerational ethics. Perhaps this is too much. Still, the issue is worth discussing; and it is worth noting that it contrasts markedly with the strong moral case for MFP with no or few reservations made by proponents of the standard argument.

To sum up, in this section I raised four basic points relevant to assessing MFP and Broome's World Climate Bank. I argued that the standard Pareto argument relies too much on the bare possibility of Pareto improvements. In doing so, it avoids taking seriously the risks posed by Pareto-violating outcomes, skewed baselines, the

extremely minimal protection afforded future generations by the Pareto criterion in the climate context, and the threat of overreach once new norms and institutions are established. Yet these should be central to any ethical analysis of intergenerational climate policy, and of intergenerational institutions more generally.

5. Optimism & Realism

The dark picture I have painted stands in sharp contrast to that presented by advocates for MFP. What explains the mismatch?

5.1. More Attractive Baselines?

I suspect that part of the answer is an optimism bias: advocates for "making the future pay" simply assume that their policies will benefit future generations against high baselines, and protect the current generation only against lower ones. Even though the standard argument is pitched in terms of Pareto improvement alone, MFP proposals are often implicitly – and occasionally explicitly – carried forward on the basis of much more demanding baselines. These pick out a subset of Pareto improvements reflecting a particular, optimistic vision of the future.

The vision goes something like this. First, the economy will continue to grow into the future once the climate threat is brought under control; second, future generations will be benefitted by that growth to such an extent that they will be much richer than the current generation is now; third, these riches ensure that the amount the future would need to pay to induce the current generation to engage in climate action does not threaten their (superior) posterity; therefore, fourth, the future can fully compensate the current generation either at low cost to itself, or at least while remaining substantially better off than the current generation, or against some other attractive baseline of prosperity. The upshot is that those attracted to the vision ultimately appeal not to the Pareto principle as such, but rather to different, more specific background ethical principles that imply that the threats to the future I discuss above are not salient after all.

Some evidence for the vision is found in Broome. On the one hand, considering future generations, he asserts:

> Common opinion among economists is that the world economy will continue to grow despite climate change, and that future people will be better off than we are. Since growth is caused by investment, this suggests that we are more than fully compensating our successors for the damage we do them through climate change. (Broome 2017, 15)

In this passage, Broome is assuming that MFP will make future generations better off than the current generation.[20]

The most plausible interpretation here, suggested by Broome's comment about investment, is that he is claiming that future generations will be much better off than people now. This interpretation is plausible because Broome is imagining continued economic growth, compounded over many decades and then centuries. Given this assumption of ongoing growth, the thought seems to be that future generations will have such a high level of welfare that whatever burdens they have to bear to cover the costs of climate mitigation will be of small account. The vision is not only "win-win" against the baseline of "business as usual" under severe climate change, but also against very high baselines of genuine benefits and absolute advantage in the future.[21]

On the other hand, this interpretation is not the only one available. In particular, Broome and Foley sometimes make major claims that appear to restrict the scope of "compensation" for the current generation. For example, at one point they claim that "leaving consumption constant ensures the present generation makes no sacrifice" (Broome and Foley, 2016).[22] Here, the focus on "leaving consumption constant" suggests that their operative assumption is that MFP will merely not decrease the consumption of the current generation; they do not say that MFP must actually increase the consumption of the current generation, and so positively benefit it.

Interestingly, this nonworsening assumption may initially appear to block worries such as mine that (i) each current generation may leverage MFP so as to enrich itself further by appropriating numerous benefits from the future, and perhaps maximally so. It may also make my more specific complaints less plausible, such as that Broome's strategy (ii) permits or encourages clawing back the gains of civilization; (iii) threatens humanity's progress; and (iv) may undermine its very survival in the longer-term.

---

[20] Notice also that in saying "*we* are more than fully compensating our successors", Broome appears to be implicitly suggesting that the current generation deserve some kind of credit (moral or causal) for the exchange. For some, this may seem uncomfortably close to an extortionist claiming credit for improving the lives of her victims by withdrawing her threats in exchange for being "paid off".

[21] Another possible baseline is more modest. Broome might be read as claiming only that future people must be at least as well off as current people, and that this functions as the relevant baseline for calculating "sacrifice". In context, I don't think that this is what he means. In any case, there are reasons to doubt that it would be a compelling baseline of "no sacrifice" from the point of view of future generations. Among other things, note that it treats whatever gains might have otherwise occurred over time as "manna from heaven" to be appropriated by the current generation, and so (again) threatens humanity's progress and the gains of civilization. (For related discussion, see Gardiner 2017.)

[22] Similarly, elsewhere they say that "we can transfer some of the [future] benefit [from climate mitigation] back in time to ourselves – enough to compensate us fully for the sacrifice we make and *leave us at least as well off as before*" (Broome and Foley 2022, 5).

Another example of restricting the claims of the current generation is when Broome and Foley sometimes claim that current owners of fossil fuels need only be compensated for the true value of their assets, and not their current market value. Their idea here is that economic theory tells us that greenhouse gas emissions are externalities that should be internalized in their price of fossil fuels so that the price reflects their true costs. Current market prices do not reflect the true price, and so are overvalued in the market. Since these fuels cannot be burned without causing catastrophic climate change, Broome and Foley say, owners of fossil fuels should be compensated for the true value of their holdings, not their present market value: "[fossil fuel investors] cannot and need not be compensated for having made a bad investment" (Broome and Foley 2022, 8). Again, this appears to be a serious constraint on the demands that can be made by the current generation.

In summary, proponents of MFP sometimes seem to presuppose more ethically attractive baselines than I have so far discussed. These baselines are based on satisfying more demanding ethical principles than potential Pareto improvement, and ones that constitute serious constraints on the current generation's pursuit of its own self-interest and are in some sense generous to future generations.

5.2. Preliminaries

Does the implicit appeal of these more demanding baselines and principles undermine my criticisms of the standard, Pareto argument? Are those criticisms now simply irrelevant? I think not. Let me begin with four preliminary points.

First, I agree that the optimistic vision opens up a world of possibility. So, (again) I will concede that versions of MFP that satisfy more attractive baselines are possible in principle. Nevertheless, we have already said that mere possibility is not enough. Less attractive outcomes of MFP are also possible, including ethically compromised or even disastrous outcomes. So, if establishing mere possibility is the point, we have not yet made much progress. Possibility is not where the action is.

Second, when we come to the action, it will be important to assess whether the economic optimism championed by Broome and others is realistic, and sufficiently so that we would be justified in relying on it. After all, there are significant reasons for skepticism. For example, [a] many will complain that it effectively assumes away the most severe threats posed by climate change, such as the possibility of abrupt shifts or runaway climate change, some of which may already be "on the cards". Similarly, [b] one might argue that the economic optimism fails to consider wider threats to human prosperity, some of which are live and accumulating (e.g., Bostrom et al.). So, for example, the optimism fails to consider environmental threats, such as those signaled in the Planetary Boundaries analysis and similar scientific assess-

ments (e.g., Rockstrom and Gaffney 2021). It also fails to mention nonenvironmental threats, such as from nuclear war (Rendall 2022).

Third, proponents of MFP need to specify which more attractive baselines they wish to endorse. There are countless possibilities, and considerable variations between them. One concern is that it is difficult to assess the plausibility of the optimism completely in the abstract. Another, vital consideration is that specifying the principle implicitly being assumed to be satisfied under optimism will likely shape concrete proposals, such as the World Climate Bank, in central ways. For instance, we might expect it to reveal critical constraints on how they operate, some of which must be incorporated at the level of institutional design. Such questions are much closer to the core issues we are interested in than discussion of mere possibility. (More on this below.)

Fourth, it seems likely that ethics will need to play a central role in making the more attractive baselines relevant after all. On the one hand, if the current generation is really to settle for leaving its own consumption constant, and to accept that future people will be much better off, how is this to be justified on the grounds of simple generational self-interest, and that alone? Alas, the whole idea seems implausible on its face. Thus, at the very least, those arguing that morality should be set aside need to explain how the more attractive baselines can be reliably achieved without ethics.

This demand seems especially pressing in the case of the two baselines mentioned above: the nonworsening constraint, and the true value of fossil fuels constraint. Why would the current generation or very powerful actors within it (such as fossil fuel companies) accept such low bars for compensation? Simple self-interest does not seem to provide an answer. On those grounds, surely the current generation as a whole will demand being made better off, and the fossil fuel barons will demand compensation that is at least close to the present market value of their holdings. Without ethics then, there seems no way forward.

By contrast, on the other hand, the more compelling positive approaches seem to be infused with ethics, not self-interest. So, for example, some might argue that members of the current generation are entitled to some level of protection against the negative effects of a climate transition, as a matter of (say) showing appropriate respect for their human rights, or ensuring they have access to central capabilities and functionings, or avoiding undue suffering. This, they would claim, flows from considerations of justice. Importantly, justice would not imply that such protection should be unlimited, nor that its key subject should be consumption, and include (say) an abundance of luxury goods. In addition, though perhaps some of this burden should be shouldered by future generations, some should presumably be borne by other parties (e.g., the rich, or those especially responsible for climate change). In

any case, the arguments would be based in ethics, and reflect the conventional debates in climate ethics, rather than the allegedly transformational approach based in self-interest.[23]

5.3. Key Questions

These preliminaries aside, let us identify some key questions. We know the "making the future pay" argument must eventually move beyond possibility. Recall that a central motivation for the approach, grounded in Pareto improvement, is practical and pragmatic. It concerns what will actually work. MFP is supposed to break the political logjam and put an end to ongoing inertia on climate mitigation. The intergenerational climate bank is the mechanism through which it is going to do this. Given this, we must ask some key questions (or clusters of questions):

(Q1) Even supposing that MFP makes meeting attractive baselines possible, what suggests that institutions and policies that really do so are actually likely to emerge from our current, morally compromised circumstances?

(Q2) What kinds of institutions would be best positioned to reliably achieve attractive baselines? What structures and safeguards would need to be put in place to facilitate their doing so?

(Q3) Is Broome's own world climate bank, based on self-interest rather than ethics, such an institution? Is it likely to reliably achieve attractive baselines? Does it embody or reflect the needed safeguards?

As far as I can tell, current discussions of MFP and Broome's WCB do not really address these questions. They effectively assume promising answers to them, rather than providing such answers. This is perhaps the optimism bias at work. Still, the door is open to enthusiasts for MFP to develop the view further by providing answers. I, for one, would welcome the effort. The possibility of meeting attractive baselines is worthy of further discussion.

Nevertheless, I must admit to thinking that the most plausible answers to the key questions are worrying. Consider first, Q1. Unfortunately, I suspect that the emergence of MFP that embodies attractive baselines for MFP for each generation is unlikely under present circumstances. This is partly because I believe that there is a

[23] Elsewhere I discuss related arguments, based in generational self-defense (e.g., Gardiner 2013, 2016; Hedahl and Fruh 2019).

governance gap. Current institutions may be good at capturing short-term concerns, especially of particular kinds (such as narrow economic concerns). But they are not good at protecting the future. Why then should we assume that pursuing MFP in such an institutional environment will actually result in better outcomes for future generations? It seems like a triumph of hope over experience.

Turning to Q2, I argue that we need genuine intergenerational institutions to fill the governance gap, and that developing them should be the subject of a global constitutional convention focused on future generations ('GCC') (Gardiner 2014, 2019). The GCC would take up the task of developing institutions that reliably fill the governance gap, and consider appropriate structures and safeguards. It would itself need to satisfy requirements of procedural and substantive legitimacy, embody some ethical ideals, and respect various constraints of justice. To me, the silence of the MFP approach on the need for more general, and distinctively intergenerational, institutional reform (beyond an intergenerational climate bank and its distinctively financial function) is a concern. Without wider intergenerational institutions, and especially the safeguards they would put in place to protect future generations, I do not see how we can have confidence in MFP as a strategy.

Third, on Q3, Broome's World Climate Bank, considered in isolation, appears to be a limited, and in various ways inappropriate, institution. It is not clear what its international and intergenerational legitimacy would be, nor why it could be relied upon to respect ethical ideals or constraints of justice. Moreover, while it could perhaps play an effective role within a more robust and intergenerationally appropriate governance regime, within the current global system it appears ripe for disfunction and also for moral corruption. Note, for example, that Broome's WCB is to be modelled on the World Bank and International Monetary Fund. These are run on the basis of financial contributions from national governments of the most powerful countries, and so largely at their behest. However, these governments have so far failed to address the climate crisis, and arguably this failure is in large part due to their short-term, narrowly economic focus. If it copies the World Bank and IMF in these respects, Broome's WCB appears poised to recreate a central dynamic that explains past political inertia as part of its essential structure.

In light of all this, should we go in a different direction? Broome's World Climate Bank proposal is justified by and conceptually linked to MFP and the "efficiency without sacrifice" (Pareto) argument. Instead, perhaps we should argue for a different kind of global banking proposal, one that asks for an intergenerational climate bank that is not essentially tied to these ideas. For instance, perhaps we should call for a Climate Justice World Bank, one that allows for contributions from future generations, in accordance to norms, principles and ideals of global and intergenerational justice. Couldn't this bank legitimately impose some burdens on the future?

Perhaps it could. After all, many approaches to climate responsibility might endorse the contribution principle ("making the future contribute" to climate action), including ones based in justice. For instance, in general, it seems plausible that most ethical approaches would allow that both future generations and the current generation should contribute something to climate mitigation as part of an intergenerational burden-sharing scheme. Indeed, as far as I know, no one has insisted that the current generation should bear all of the burdens of climate mitigation, no matter what the wider circumstances. On the contrary, it often seems plausible that some burden-sharing across generations is appropriate. One reason that this approach seems reasonable for climate is that at this point in history the climate problem has been evolving over many decades, so that one thing that needs to be addressed is how to share the burdens created by the moral failures of past generations (e.g., those in charge in the 1960-1989 or 1990-2019). Thus, there is perhaps a presumption that just intergenerational institutions, including a Climate Justice World Bank, would allow for some intergenerational burden-sharing.

Nevertheless, we would need to be very clear that the call for a Climate Justice World Bank is a very different proposal than Broome's call for his Future-Saddling, Present-Sparing World Climate Bank. Consider four initial ideas. First, the Climate Justice World Bank is motivated by ethics and constrained by justice, rather than driven by self-interest. Second, since it aims to reflect norms, principles and ideals of justice, the design of the CJWB should not be left solely, or even primarily to economists. Instead, it should draw on the expertise of disciplines and professions centrally tied to concepts of justice. Third, I would suggest that the CJWB should be specifically constructed so as to combat the central threats of intergenerational buck-passing, the tyranny of the contemporary and moral corruption. Fourth, I also believe that any call for a climate justice world bank is probably best seen as part of a wider institutional proposal to be considered by the global constitutional convention for future generations, rather than a standalone institution. In summary, I conclude that the call for a Climate Justice World Bank is one with a strikingly different flavor to Broome's call for a Future-Saddling, Present-Sparing World Climate Bank.

6. Conclusion

Where does this leave us? I hope to have made clear that neither "making the future pay" in general, nor the Pareto argument and proposal for a Broomean World Climate Bank more specifically, are compelling considered in the abstract or in isolation. While it is possible to imagine attractive versions of each, it is equally possible to envision deeply ethically compromised, and even horrifying scenarios.

Fortunately, possibility should not be our standard; instead, we must be more realistic and look at the root of our problems.

Sadly, this creates fresh concerns for the "making the future absorb [the burdens]" strategy. When we look at the context of international climate policy, and in particular its history of failure, and when we note the roles of current national and international institutions in that failure, there is little reason to believe that morally attractive versions of an intergenerational climate bank are likely to emerge from them. Indeed, plausibly, in our current institutional context Broome's proposal is more likely to reproduce the problems we face than to resolve them. It may also amplify such problems.

More generally, I suspect that progress depends on developing better institutions, including by filling the governance gap for future generations, and incorporating appropriate safeguards and protections against intergenerational buck-passing and moral corruption. Elsewhere I propose that a global constitutional convention for future generations is the appropriate forum for developing such reforms. Perhaps that would be the best context in which to discuss an intergenerational climate bank and "making the future contribute" to climate finance more generally. Notably, the GCC would put ethics at its heart, in its core motivation, principles and constraints. It would resist the appeal to simple self-interest that motivates many rival approaches, including Broome's World Climate Bank. One upshot is that, while a Climate Justice World Bank may be desirable, there is a strong presumption that it would look very different from Broome's World Climate Bank.

References

Arrhenius, Gustaf. 2022. "A World Climate Bank: A Response". *Global Challenges Foundation*, April.

Broome, John. 2012. *Climate Matters: Ethics in a Warming World*. New York: Norton.

-----. 2017. 'Do Not Ask for Morality'. In A. Walsh, S. Hormio, & D. Purves, (Eds.), *The Ethical Underpinnings of Climate Economics*. Routledge.

Broome, John, and Duncan K. Foley. 2016. "A World Climate Bank." In *Institutions for Future Generations*, edited by Inigo Gonzalez-Ricoy and Axel Gosseries, 156–169. Oxford: Oxford University Press.

Broome, John, and Duncan K. Foley. 2022. "A World Climate Bank". *Global Challenges Foundation*, April.

Gardiner, Stephen M. 2001. "The Real Tragedy of the Commons." *Philosophy and Public Affairs* 30.4: 387–416.

_____. 2003. "The Pure Intergenerational Problem." *The Monist* 86.3: 481–500.

_____. 2004. "The Global Warming Tragedy and the Dangerous Illusion of the Kyoto Protocol." *Ethics and International Affairs* 18.1: 23–39.

_____. 2006. "A Perfect Moral Storm: Climate Change, Intergenerational Ethics and the Problem of Moral Corruption." *Environmental Values* 15: 397–413.

_____. 2011. *A Perfect Moral Storm: The Ethical Tragedy of Climate Change*. Oxford: Oxford University Press.

____. 2013. "Why Geoengineering is Not a Global Public Good, and Why it is Ethically Misleading to Frame it as One." *Climatic Change* 121: 513–525.

_____. 2014. "A Call for a Global Constitutional Convention focused on Future Generations." *Ethics and International Affairs* 28.3: 299–315.

_____. 2017. "The Threat of Intergenerational Extortion." *Canadian Journal of Philosophy* 47.2–3: 368–394.

_____. 2019. "Motivating (or Baby-Stepping Toward) a Global Constitutional Convention for Future Generations." *Environmental Ethics* 41: 199–220.

_____. 2021a. "Intergenerational Climate Extortion." *Rivista di Filosofia del Diritto* (Italian Review of Legal Philosophy).

_____. 2021b. "Debating Climate Ethics Revisited." *Ethics, Policy and the Environment* 24.2, 89–111.

_____. 2022a. "Is the Paris Climate Agreement Another Dangerous Illusion?" In *The Norwegian Academy of Science and Letters Yearbook 2021*. Oslo: Novus forlag, 2022.

____. 2022b. "On the Scope of Institutions for Future Generations". *Ethics and International Affairs* 36.2, 157–178.

_____. In press. "Climate Change and the Intergenerational Arms Race." In *Environmental Ethics for Canadians*, ed. B. Williston. 3rd ed. Oxford: Oxford University Press.

Gardiner, Stephen M. and David Weisbach. 2016. *Debating Climate Ethics*. Oxford: Oxford University Press.

Hedahl, Marcus and Fruh, Kyle. 2019. "Climate Change as Unjust War." *Southern Journal of Philosophy* 57.3: 378–401.

Intergovernmental Panel on Climate Change (IPCC). 2021. *Climate Change 2021: The Physical Science Basis.* Cambridge: Cambridge University Press.

Maltais, Aaron. 2015. "Making Our Grandchildren Pay for Mitigation." In *The Ethics of Climate Governance,* edited by Catriona McKinnon and Aaron Maltais, 91–109. Lanham: Rowman and Littlefield.

Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity*. New York: Hachette.

Parfit, Derek. 1985. *Reasons and Persons*. Oxford: Oxford University Press.

Rendall, Matthew. 2011. "Climate Change and the Threat of Disaster: The Moral Case for Taking Out Insurance at Our Grandchildren's Expense." *Political Studies* 59: 884–89.

Rendall, Matthew. 2021. "Public Debt and Intergenerational Ethics: How to Fund a Clean Technology 'Apollo Program'?" *Climate Policy* 21.7: 976–82.

Rockstrom, J., and O. Gaffney. 2021. *Breaking Boundaries: The Science Behind Our Planet.* London: DK Publishing.

Katie Steele[1]

Longtermism and Neutrality about More Lives

Longtermism is associated with the claim that we have moral reason to attempt to reduce the risk of 'futuristic threats' to humanity's survival, even taking into account considerable opportunity costs for present people. A tempting response to this uncomfortable claim is to pin it to a specific way of conceiving moral reasons directed at human welfare, namely a *totalist* one that turns on more worthwhile lives being inherently better. The opposing conception is that more worthwhile lives is neither better nor worse, but rather inherently *neutral*. The well-known 'greediness of neutrality' (Broome 2004, 2005), however, has the wider implication that conclusions about futuristic threats of any kind are not greatly dependent on *totalist* moral mathematics. In deliberating about the goodness of actions that affect present persons, our predictions about those who may or may not exist later inevitably matter.

[1] Australian National University, and Institute for Futures Studies; katie.steele@anu.edu.au

1. Introduction

The momentum of the *Longtermism* movement—escalated by the publication of Will MacAskill's (2022) book *What We Owe the Future*—owes in part to an exciting vision of the power of the present generation to give humanity a leg up towards a brighter, longer-lasting future. The headline calls to action are for significant investments to reduce the risk of what are perceived as 'futuristic threats', e.g., premature human extinction due to engineered pathogens, or a long-enduring tyranny enabled by artificial superintelligence (ASI). The idea is that these are the threats to human flourishing that, while not necessarily immediate, span the longest time scales and are most impactful. Indeed, the time scales are so long that we have very strong moral reasons to pursue even small reductions in the risk of these threats materialising.

While a gripping vision, many find the idea of devoting significant resources to addressing future threats deeply troubling if it means diverting resources from, say, tackling the more immediate and cumulative problem of climate change, or reforming existing institutions to reduce economic inequalities and other injustices. I too find this troubling. Many simply dismiss the longtermists' conclusions—or at least those concerning premature human extinction—on the grounds that they rely on a faulty premise regarding the fundamental nature of moral reasons directed at human welfare. The supposed error lies in treating more worthwhile lives as inherently better since this increases *total* positive human welfare. The proposed alternative is that more worthwhile lives is neither better nor worse, but rather inherently *neutral*. Kieran Setiya[2], for example, has made this claim as follows:

> But if neutrality is right, the longtermist's mathematics rest on a mistake: the extra lives don't make the world a better place, all by themselves. Our ethical equations are not swamped by small risks of extinction. And while we may be doing much less than we should to address the risk of a lethal pandemic, value lock-in, or nuclear war, the truth is much closer to common sense than MacAskill would have us believe. We should care about making the lives of those who will exist better, or about the fate of those who will be worse off, not about the number of good lives there will be.

One problem with this statement is that it groups different kinds of future threats together, both those that affect the number of worthwhile lives (premature human extinction) and those that do not affect the number of lives at all, just the quality of these lives ('value lock-in', or a long-enduring tyranny). But consider just the former

[2] This is from his (2022) article in *Boston Review*.

kind of threat, for which the question of neutrality is pertinent. If the diagnosis stated above were right, we would not need to engage with the longtermists' empirical premises regarding premature human extinction, given various courses of action. Nor need we consider how they propose to manage uncertainty about such consequences in decision making.

But unfortunately this short circuit strategy does not work. I will argue that none of the longtermists' prominent conclusions rest precariously on a deeply divisive view about the fundamental nature of moral reasons directed at human welfare. That includes, somewhat surprisingly, the conclusion that we have strong moral reason to reduce the risk of premature human extinction (or at least conclusions in this ballpark). I show this by revealing how *neutrality* about additional worthwhile human lives does not preclude longtermist concerns, including premature human extinction, that are associated with *totalism* about additional worthwhile human lives. It is not just that there are welfarist reasons, whether totalist or neutral, to care about reductions in the risk of *specific* kinds of future threats—in particular those which affect the welfare of a population of fixed size.[3] Rather, there are welfarist reasons, whether totalist or neutral, to care about reductions in the risk of *many* kinds of future threats—including those which determine the size of the population. Moreover, I will argue that the neutral approach may lead to problematic *non-longtermist* conclusions in cases where additional lives are worthwhile and yet fall short of a utopian level of welfare.

This does not mean that the longtermists' calls for action are well founded. To effectively challenge them, however, we must instead redirect attention to the *non-moral* premises on which their robust appeal—given less fundamental disagreements about welfare—depends. It is one thing to care about the *possibility* of a reduction in the risk of a future threat. It is quite another to take there to be strong moral reason to *actually* pursue, in a direct way, a reduction in the risk of that threat. This point may be clear enough in the case of future threats to a fixed population. I therefore start out—in the next section—with this case. We see that the empirical premises are crucial to the choice conclusions. With select premises, differing accounts of moral reasons directed at human welfare, whether totalist or neutral, converge on seemingly radical choice conclusions. The remainder of the chapter argues that the case of reducing the risk of premature extinction is not much differ-

[3] That is the line pursued by Hilary Greaves and Will MacAskill (2021, p. 18) in their brief discussion of the sensitivity of longtermist claims to the details of moral theory. They say, cautiously: 'According to the spirit of a person-affecting approach, premature extinction is in itself at worst neutral: if humanity goes prematurely extinct, then there does not exist any person who is worse off as a result of that extinction, and, according to a person-affecting principle, it follows that the resulting state of affairs is not worse. The far-future benefits of extinction risk mitigation may therefore beat the best near-future benefits only conditional on controversial population axiologies.'

ent; seemingly radical choice conclusions to this end are not greatly sensitive to the difference between totalist and neutral approaches to moral reasons directed at human welfare. That is, here too, totalism in the longtermists' mathematics plays a relatively minor role.

2. Future threats to a fixed population

First some set up that will assist the discussion throughout the paper. We are concerned with the deliberations of a decision maker who faces choices with far-reaching implications. This may be a single individual or a governing body. In line with the orientation of longtermists, we will consider just the moral reasons directed at general human welfare—what might be called reasons of *impartial beneficence*, but which we will refer to simply as 'welfarist reasons'—that bear on this decision maker's choice. There may be other moral reasons, and perhaps more important ones, that bear on her choice. Perhaps it is reasonable to assume that all else is equal with respect to these other moral reasons, which might concern special relationships or rights and obligations. But perhaps that is not a reasonable assumption for the kinds of applications the longtermist has in mind. Our decision maker may in any case have further non-moral reasons for choice, for instance, personal or prudential reasons. There is an important and overarching question of how *all* her reasons, both moral and non-moral, may be weighed against each other. But we put this overarching question aside here. Our discussion is limited to the welfarist reasons that bear on a decision maker's choice, as this is arguably how the longtermists' claims or 'choice conclusions' are best interpreted.

Consider the claim that our decision maker has strong welfarist reasons to pursue a 'far-sighted' option pertaining to the use of some resource X that would exclusively benefit the many who will exist in the further future. For instance, the option in question might involve spending X on reducing the risk of debilitating 'value lock-in' or long-lasting tyranny enabled, say, by artificial superintelligence (ASI). That our decision maker has strong welfarist reasons to pursue this option depends on a number of premises, regarding: i) the empirical decision set-up, ii) the proper resolution of uncertainty and iii) the nature of the welfarist reasons. Note that one can think of the welfarist reasons as dependent on the relative moral goodness of the outcomes with respect to human welfare (or the 'welfarist good' for short).[4]

[4] One can interpret 'welfarist good' in a very general way; the associated ranking of outcomes may, for instance, be 'context sensitive' such that *transitivity* is not satisfied across choice sets. That said, I will later assume transitivity in my discussion of the implications of neutrality in Section 4.

Table 1: Choice Problem for a Fixed-Sized Population

	State 1		State 2	
	near	**far**	**near**	**far**
short-sighted	modest	none	modest	none
far-sighted	none	extreme	none	none

Table 1 exemplifies the substantive nature of the first kind of premises. The table depicts just two competing options for spending X, or two deviations from the status quo (which is not explicitly represented, but is the reference point for describing the outcomes of the other options). All of the possible outcomes contain populations of the same size.[5] The outcome of the 'far-sighted' option, unlike the 'short-sighted' option, depends on which of the possible states of the world turns out to be actual. The first state in the table, 'State 1', is highlighted to emphasise that the welfarist comparison of the options depends primarily on the outcomes under this state. This is the state of the world in which pursuing the 'far-sighted' option makes a significant difference to the welfare of the 'far' group, which contains a relatively enormous number of people further away in time—in aggregate, this is an extremely positive change in welfare from the status quo. Even if we assume that State 1 is extremely unlikely compared to State 2, the extraordinary welfarist difference in the outcomes under State 1, let us say, is such that the relatively minor welfarist difference in the outcomes under State 2 is not important.

This brings us to the question of what our decision maker has most welfarist reason to do. If it is assumed that she faces the choice problem represented by Table 1, then it is not such a large jump to the conclusion that she has (even strong) welfarist reason to pursue the 'far-sighted' option, given standard ways of resolving uncertainty.[6] This conclusion does not depend at all on the difference between the

[5] We might add an even stronger premise that all of the possible outcomes contain populations constituted by the same people. Then the choice recommended by welfarist reasons would be even less sensitive to differences in the various accounts of the welfarist good. But since the focus here is the fundamental difference between totalist and neutral approaches to the value of additional worthwhile lives, we will put aside more extreme 'person affecting' versions of the latter, whereby the relative goodness of a pair of outcomes depends only on their welfare effects with respect to those who exist on both outcomes.

[6] While we do not have the space to properly address the resolution of uncertainty in this setting, Table 1 facilitates an ex post approach, which involves evaluating the respective outcomes and then using this information, as well as the probability of the outcomes, in evaluating the options. A 'standard way' of evaluating the options would be one that conforms to (some conservative generalisation of) the expected value principle, whereby the higher the expected value the better the option.

totalist and neutral approaches to the value of additional worthwhile human lives, given that the size of the population is not in question. Just *how robust* is the conclusion that our decision maker has most welfarist reason to pursue the far-sighted option depends on some further considerations that are elided in Table 1. For instance, it will be more robust if the many further away in time are the worst off and the welfare increments for each of these individuals in State 1 is moreover large enough for these to count as 'relevant claims'[7] when compared to the welfare increments at stake for each of the individuals nearby in time. So there is some slightly more detailed specification of Table 1 such that, according to a large range of views about the relative welfarist goodness of outcomes, our decision maker has (even strong) welfarist reason to pursue the far-sighted option.

The argument that we have strong welfarist reason to pursue a reduction in the risk of a long-enduring tyrannical regime enabled by ASI is most puzzling and compelling when presented along the lines of Table 1. It is puzzling because it is counter to ordinary practice to think that our welfarist reasons direct us to focus on the far future. And yet it is compelling because the conclusion seems irresistible since robust to differences in views about the welfarist good. But the conclusion is only irresistible on the assumption that the decision problem really looks something like Table 1. This seems initially plausible, if we consider the possibility, however remote, of a long-enduring tyrannical regime. The aggregate welfare benefit of reducing the risk of this outcome, even very slightly, is presumably large enough to easily outweigh any modest welfare benefits that would accrue to relatively few people nearby in time. But a decision problem like that in Table 1 depends on further empirical premises: that acting to directly reduce the risk of specific future threats will really work, and that the opportunity costs, or the relative benefits of alternative actions, are located just within the nearby time period and are thus relatively small.

In the next section, I claim that the same sorts of considerations apply to longtermists' arguments that we have strong welfarist reasons to pursue a reduction in the risk of premature human extinction. Here it may seem that the empirical details matter less; that one can get off the longtermists' wagon simply by denying the totalist approach to the value of additional worthwhile human lives. But we will see that this is a mistake.

3. Future threats to population size

Let's for the moment assume a totalist approach to the value of additional human lives. On such an approach, every extra life with positive welfare—assuming welfare

[7] For an influential welfarist view that turns on aggregating only 'relevant claims', see Voorhoeve (2014).

is scaled so that a life of zero welfare is neither good nor bad for the person living it—increases the welfarist good of an outcome, and every extra life with negative welfare decreases the welfarist good of an outcome. (For the specific totalist approach known as *total utilitarianism*, for instance, the overall welfarist good of an outcome is just the sum of the welfare of all who exist in that outcome.) It is not hard to see that, on a totalist account of the welfarist good, greatly increasing the number of persons with positive welfare who will exist is, all else equal, a very good thing to do.

Accordingly, it seems prima facie plausible that we have strong welfarist reasons, if some version of totalism is right, to reduce the risk of premature human extinction. After all, premature human extinction cuts off an indefinitely long period of human history that plausibly contains on balance extraordinary amounts of positive welfare. Just like in the case of a long-enduring tyranny, the welfare benefit of reducing the risk of premature human extinction, even very slightly, seems then large enough to easily outweigh the welfare benefit of any shorter-term project.

Table 2: Choice Problem for a Varying-Sized Population

	State 1		State 2	
	necessary	**possible**	**necessary**	**possible**
short-sighted	modest	none	modest	none
far-sighted	none	extreme	none	none

That is, it seems that our plight as decision makers may plausibly resemble Table 2, where the 'far-sighted option' is to spend some resource X on directly trying to reduce the risk of premature human extinction. And if so, we would have strong welfarist reasons, if these reasons are broadly totalist at least, to pursue this option. As before, the table depicts the outcomes for two options or deviations from the status quo. The outcomes are presented with respect to a partition of the possible overall population: in this case, the first group contains some inevitable number of people, the 'necessary'-sized group, and the second group contains some further number of people that is contingent on the natural and agential forces pertinent to this choice problem, i.e., the 'possible'-sized group.[8] Again, 'State 1' is highlighted. This is the state of the world in which pursuing the 'far-sighted' option makes a

[8] This is a variant of the 'necessary people' versus 'possible people' distinction. There is a subtle difference in that here we are focusing just on the size of the population. By contrast, the 'necessary people', for instance, are the specific people who will exist, whatever choice one takes; they are not merely those who make up some fixed number of people who will exist, whatever choice one makes.

significant difference to the welfare of the 'possible' group, because it adds an enormous number of people to this group.

Many are unmoved, however, by the use of Table 2, or something like it, to argue that there are strong welfarist reasons to directly aim at reducing the risk of premature human extinction. The thought is that any such argument is a non-starter since it depends precariously on welfarist reasons of a totalist kind. The alternative approach to welfarist reasons that many find attractive, especially in view of future threats that turn on premature human extinction, is what was referred to above as the neutral approach, or 'neutrality' for short. On this approach, extra lives with positive welfare are neither good nor bad in and of themselves, but are rather neutral.[9] It does seem that welfarist reasons of a neutral kind would provide no reason to aim at ensuring the existence of more people, especially when there are significant opportunity costs to presently existing people, or more generally to a population of guaranteed size. Only under totalist approaches is there an extraordinary difference in welfare between the two outcomes under 'State 1', a difference that effectively determines the welfarist comparison of the options even if 'State 1' has extremely small probability.

Before going on in the next section to examine the implications of neutrality, I will pause to state the position more carefully. The thought is that there is some range of welfare that we can refer to as the *neutral range*, whereby adding an extra person whose welfare falls within that range does not make the world better or worse in and of itself. The neutral range is typically thought to have a lower bound equating to the value of a life that is neither good nor bad (zero, by convention). Lives with welfare below the neutral range—those that are bad for the person living them—are thought to detract from the welfarist goodness of an outcome.[10] The upper bound of the neutral range could be very high or even infinite. Arguably the common thought is that any good life, no matter how good, does not in itself make the world a better place. Wlodek Rabinowicz (2022, 116) refers to this as the 'radical' interpretation of neutrality, whereby the neutral range of welfare extends from zero all the way up to the value of a maximally good life, or else infinity if there is no upper bound on a person's welfare. (He contrasts this with the 'moderate' interpretation, whereby the neutral range of welfare extends from zero to some positive, not-too-high level of welfare.) Here we will initially stick with the 'radical' interpretation,

[9] Not everyone thinks it is obviously a mistake to worry about premature human extinction, including some who endorse neutrality. For instance, Frick's (2017) project is to offer an account, consistent with neutrality, of our reasons to prevent premature human extinction.

[10] That is, neutrality is typically spelled out in a way that is sensitive to the so-called 'Procreation Asymmetry', whereby there is a duty not to bring into existence a person with a bad life, presumably because this would be bad, and yet there is no duty to bring into existence a person with a good life, presumably because this is neither good nor bad.

since, in spite of Rabinowicz's label, that is arguably what many advocates of neutrality have in mind.

Neutrality seems initially to have much to recommend it. Indeed, Jacob Nebel (2019) and Johann Frick (2020) have provided rich accounts of the notion of value that underpins intuitions supporting neutrality. The idea is that positive welfare is not something we have *unconditional* reason to bring about. Nebel and Frick propose instead welfare-related reasons that are *conditional* on the person in question's existence, or the *bearer* of welfare. This leads to an account whereby outcomes are ranked according to a notion of conditional welfare value: It is better that bearers of welfare have positive rather than negative welfare (and more generally higher rather than lower welfare), and it is better for there to be no bearer of welfare than for there to be one with negative welfare, but it is neither better nor worse for there to be no bearer of welfare than for there to be one with positive welfare. Surely then, one might think, preventing premature human extinction (or rather, increasing the number of lives with positive welfare) is not worth any sacrifice if one assumes neutrality.

4. Implications of Neutrality

The implications of neutrality turn out to be difficult to glean, however, just by contemplating the *kind* of welfare value that underpins this approach. John Broome (2004, 2005) showed that the resulting ranking of outcomes with respect to welfare (or 'welfarist ranking') is more complicated than first appearances suggest. Here I will extend Broome's insights to outcomes with long time horizons that allow for much variation in the number of persons who exist. We will see that neutrality does not in fact preclude acting on welfarist reasons to directly reduce the risk of premature human extinction, even at great opportunity cost to those living in the present, or more generally to the fixed-size population that is bound to obtain. In fact, neutrality does not preclude even more counter-intuitive conclusions.

4.1 Greedy Neutrality

The starting point for Broome (2004, 2005) is what precisely is the relationship, in terms of the welfarist good, between some original population and that population with an additional life of positive welfare. Neutrality says that the former, with the additional life, is neither better nor worse than the latter. Should this be interpreted as indifference? Arguably not, since if the indifference and better-than relations satisfy *transitivity*, we get an inconsistency with the *(strong) Pareto principle*. Consider the following case, where the population outcomes, *A*, *B*, and *C*, are represent-

ed as vectors. Each entry in the vector corresponds to a particular person and gives their lifetime welfare, with '-' representing that the person does not exist on that outcome:

$A = (2, 3, -)$

$B = (2, 3, 4)$

$C = (2, 3, 5)$

On the indifference interpretation, neutrality here implies indifference between A and B ($A = B$) and between A and C ($A = C$). But then by transitivity of indifference, we would get $B = C$. But C weakly Pareto dominates B, so by the (strong) Pareto principle C is better than B ($C \succ B$). Hence, an inconsistency.

Broome (2004, 2005) proposes instead that neutrality be fleshed out in terms of the relation of incommensurability. In that way, neutrality conforms with both strong Pareto and transitivity of the better-than relation with respect to the welfarist good. For our example, A is then incommensurable with both B and C ($A \approx B$ and $A \approx C$). Since incommensurability is not transitive, we may yet have $C \succ B$. Indeed, the classic examples used to illustrate the difference between incommensurability and indifference have precisely this form. For instance, in the ranking of holiday destinations, to say that a rugged wilderness escape, A', and a cultural city sojourn, B', are incommensurable (that is, $A' \approx B'$) is to say that there is some enhancement (say a fixed amount of extra spending money) to the cultural city sojourn (yielding C', where $C' \succ B'$), such that A' and C' are also incommensurable (that is, $A' \approx C'$).

What Broome further draws attention to—our starting point for the next section—is that the incommensurability of neutrality is 'greedy'. It allows good and bad changes to the original population to be offset by the addition of lives with positive welfare to the original population. (What is meant by 'offset' here is that, on net, there are no welfare reasons for or against the new augmented population involving gains or losses to the original population, as compared to the original population; since the two are incommensurable.) A simple example (owing to Rabinowicz 2009) will suffice to demonstrate both aspects of this point:

$D = (3, 4, -)$

$E = (3, 4, 1)$

$F = (3, 3, 3)$

$G = (3, 3, -)$

According to neutrality, D and E are incommensurable with respect to the welfarist good ($D \approx E$) and similarly F and G are incommensurable ($F \approx G$). But plausibly $F \succ E$, since F has greater equality, and also greater total welfare. But since E is not worse than D, F cannot be worse than D. And yet the move from D to F involves a decrease in welfare for someone and the addition of a new person with positive welfare. So we see that neutrality entails that the addition of a person with positive welfare (in the neutral range) can offset a decrease in welfare for those in the original population. (Broome says rhetorically that neutrality about added lives can 'swallow up' the original people's loss in welfare, neutralising it, and making the change overall not bad.)

Now consider G and E. Since, according to neutrality, F is not better than G, and F is itself better than E, then E cannot be better than G. But the move from G to E involves an increase in welfare for someone and the addition of a new person with positive welfare. So we see that neutrality entails that the addition of a person with positive welfare (in the neutral range) can offset an increase in welfare for those in the original population. (Again, we might say that neutrality about added lives can 'swallow up' the original people's gains, neutralising it, and making the change overall not good.)

Broome (2004, 2005) claims that the appeal of neutrality depends on there being no swallowing up of this sort, and so the incommensurability interpretation of neutrality shows this approach to be based on some false hope. Others disagree (e.g., Rabinowicz 2009, Frick 2017); they suggest that greediness is not necessarily an unwelcome consequence of neutrality for those who find it appealing. In any case, the focus here is to simply extend Broome's investigations of the greediness of neutrality to the long-term or large-scale setting where many lives may be at stake. We will ultimately compare, in this setting, the neutral and totalist approaches to the welfarist good.

Note that our discussion will treat transitivity as a constraint on the better-than relation with respect to the welfarist good. (Further assumptions will also be introduced for ease of explication, more on which below.) On a broader notion of the welfarist good—one that simply tracks reasons for choice in any given choice context—the associated 'more-choice-worthy-than' relation need not be transitive across choice contexts. For instance, Frick (2022) allows that the choice-worthiness relation between pairs of options may change depending on the context: for the above options, E may be no less choice-worthy than D in a pairwise comparison, but less choice-worthy where the option set also includes F, because in that case the choice of E would involve unnecessary inequality (compared to F). This sort of context dependence *may* make neutrality less greedy overall, but it depends on the details of the fully formulated account. Teruji Thomas (forthcoming, sections 5.1 &

5.2) offers fully formulated accounts along these lines that he claims are defensible in the context of uncertainty, but, on these accounts, as per my discussion below, the lives of additional people matter in the welfarist comparison of options. So I tentatively suggest that weakening the transitivity assumption would not dramatically change the overall story in what follows in the ways that defenders of neutrality who wish to resist longtermist conclusions might hope.

4.2 Greedy Neutrality and Longtermism

Let us turn then to the long term or large scale. Even if the greediness of neutrality is not particularly surprising or unwelcome on a small scale, things may look different on a large scale. The investigation of the latter setting is, however, not so straightforward. We saw that simple demonstrations of the greediness of neutrality, like that above, depend on very few assumptions regarding the welfarist ranking of options. When it comes to settings in which many lives are at stake—including decision problems for which the choice of option affects the threat of premature human extinction—we must introduce further assumptions to assess the implications of neutrality. Indeed, it helps to work with a rather specific account of the welfarist good and leave it largely to the reader to consider how the results would change for nearby accounts. At the very least, we can say that the results demonstrated here are not *precluded* by neutrality.

For the sake of clear explication, assume then that the welfarist good has the following features: i) populations of differing size are compared in a way that conforms to neutrality, whereby adding lives with positive welfare to an original population yields an augmented population that is incommensurable with that original population, and ii) populations of the same size are compared in terms of average welfare, the higher the better.[11] (The second feature is clearly arbitrary in the context of our discussion; it might just as well be some other approach to comparing populations of the same size.)

This notion of the welfarist good makes clear that neutrality does not commit one to a view whereby populations of the same size constituted by different people are incommensurable. On the account just outlined, such populations are ranked according to their average welfare. So the following statement from MacAskill (2022, 175) is rather misleading:

[11] This may well be an alternative description, or at least a partial description, of a specific version of 'critical-range' utilitarianism (see Rabinowicz e.g., 2009, 2022), in which the 'critical range' extends from zero (representing the welfare of a life that is neither good nor bad for the person living it) to positive infinity.

> Consider two people, Alice and Bob. If we keep fossil fuel subsidies, Alice will be born in 2070. If we end fossil fuel subsidies, Alice will not be born and Bob will be born instead. Both have happy lives, but, because climate change will be less extreme without fossil fuel subsidies, Bob will be happier than Alice would have been. According to the intuition of neutrality, we do not have reason to ensure that Bob exists rather than Alice. According to the intuition of neutrality, preventing Alice's existence is neither good nor bad, and bringing Bob into existence is also neither good nor bad. So doing both at once is neither good nor bad.

MacAskill describes a violation of what Derek Parfit (1984, 367) calls the 'No Difference View', since the change in identity here between the happy and happier person *does* affect the comparison of the two outcomes. But neutrality does not itself imply violations of No Difference. In fact, defenders of neutrality typically also defend No Difference (see, e.g., Frick 2020). For the example above, that would mean that ending fossil fuel subsidies is the better option, since were it not for a change in identity from Alice to Bob, this option weakly Pareto dominates the other. The change in identity does not affect this ranking.

So, neutrality does not lead to so much incommensurability in the comparison of options as MacAskill suggests. But it still leads to a lot of troubling incommensurability owing to the greediness of neutrality.

Consider the case where added lives 'swallow up' suffering or bad changes for the original population. The recipe for generating such cases of swallowing up is as follows: Take an 'original population'. Add n lives right at the cusp of being worth living, or just above zero welfare. Call this the 'augmented population'. It is incommensurable with the original population with respect to the welfarist good. Now there will be various populations of the same size as the augmented population that are better than that population. These '+n-populations' are thus not worse than the original population; they are either incommensurable with or better than the original population. But some of these +n-populations will involve welfare losses for the original population. These are the populations in which the added lives 'swallow up' losses to the original population. For our account of the welfarist good in which same-sized populations are compared in terms of their average welfare, the +n-populations will include populations with any given sum of welfare loss to the original population; this loss must be effectively compensated by at least as great a sum welfare gain for the added n people, compared to what they each could have had in the augmented population, which was close to zero welfare.

The swallowing up gets even worse. It is not just that any welfare loss to the original population can be swallowed up by sufficient total gains for the added people relative to their lives being only just worth living (or close to zero welfare). Worse,

the more people added to the original population, or the greater that *n* is, the less their respective welfare levels need to surpass zero for their sum gains to be sufficient.

Return then to the choice problem in Table 2, where under State 1, the 'far-sighted' option results in many more lives with positive welfare. On a totalist approach to the welfarist good, the 'far-sighted' option is better than the 'short-sighed' option, provided the difference in the goodness of the outcomes under State 1 is sufficiently large relative to the (very small) probability that State 1 is true. Can we get around this uncomfortable conclusion with a neutral approach to the welfarist good? Neutrality does not preclude the 'far-sighted' option being incommensurable with the 'short-sighted' option. This is to say there is no positive welfarist reason in favour of present sacrifices (forgoing present welfare gains) to pursue additional future lives with positive welfare. But equally, there is no welfarist reason *against* present sacrifices to pursue additional future lives. Acting to reduce the risk of premature human extinction may therefore not have less merit, on welfare grounds, than acting to mitigate present suffering. That would be so, at least, if the choice problem looked somewhat like that described in Table 2.

4.3 Greedy Neutrality and Non-Longtermism

We see that neutrality does not preclude choice conclusions that depend on the number of people who will live and the quality of their lives: great gains in worthwhile future lives may be worth (or can at least *offset*) sacrifices in welfare for those in the present. That should already give defenders of neutrality pause—this approach to welfare does not insulate present decision making from considerations of varying population size, at least not to the extent that one might have hoped. It gets worse, however. Neutrality does not preclude further disturbing implications that are not shared by totalist approaches. We will see that not only are the advantages of the neutral compared to the total approach to welfarist reasons less pronounced than first appearances suggest, but the former also has marked disadvantages.[12]

To see this, let us consider now the other side of greedy neutrality: cases in which welfare *gains* to the present are swallowed up by the addition of future lives with positive welfare. The recipe for generating such cases of swallowing up is similar to

[12] This is not surprising in light of the 'impossibility theorems' of population ethics (see, for instance, Arrhenius ms.). Standard forms of totalism notoriously imply the 'repugnant conclusion'. Alternatives designed to avoid that conclusion have other problems, for instance they imply the 'sadistic conclusion'. Neutrality is somewhere in the middle. It may lead to a less repugnant conclusion, but the other side of the coin is that it leads to a less, say, sadistic conclusion. In other words, we might expect that a neutral approach to the welfarist good will not have the extreme counterintuitive properties of other accounts, but will have many counterintuitive properties nonetheless.

that outlined above: Take an 'original population'. This time add n lives with extremely high welfare; some finite level of welfare will work in the recipe since we are assuming that the neutral range is unbounded. Again, call this the 'augmented population'. It is, by construction, incommensurable with the original population with respect to the welfarist good. Now there will be various populations of the same size as the augmented population that are worse than that population. These '+n-populations' are thus *not better* than the original population; they are either incommensurable with or worse than the original population. But some of these +n-populations will involve welfare gains for the original population. All that is required for a population to qualify as a +n-population is that the sum welfare gains for the original population are outweighed by at least as much sum welfare loss for the added n people, compared to their welfare in the augmented population (i.e., extremely high welfare). These are the populations in which the added lives 'swallow up' gains to the original population.

This is a very troubling implication for neutrality, at least when formulated with the features specified above.[13] Note that the first feature is a neutral range that is unbounded from above, extending to positive infinity. It is arguably that first feature, rather than the second which specifies an average welfare approach to comparing populations of the same fixed size, which yields the most trouble. When the neutral range is unbounded from above, the added lives to the original population *could have had any positive welfare whatsoever* and the resulting augmented population would still count as incommensurable with the original population. So whatever the *actual welfare* of the added lives, a sum welfare loss of any size whatsoever is incurred, relative to the augmented population. This loss can outweigh any sum welfare gain to the original population. In short, neutrality does not preclude a very severe kind of swallowing up of gains to an original population. We see that adding *any number of lives to an original population*, at *any positive level of welfare*, 'swallows up' *any sized welfare gain* to the original population. That is, the resulting population is *no better* (incommensurable, or even worse) than the original population.

We can refer to this as a (worrying) *non-longtermist* implication of neutrality: increasing the size of the human population, even if this is a *win-win* scenario with no intertemporal trade-offs, is no better than the status quo. So, for instance, the mitigation of climate change, or fantastic advances in medicine, insofar as *they increase the size of the human population by adding worthwhile lives*, are not better than the status quo, even if they also come with many welfare benefits for those in

[13] It amounts to a violation of an axiom known as *Dominance Addition* which Gustaf Arrhenius (ms., 307) articulates as follows: 'An addition of lives with positive welfare and an increase in the welfare of the rest of the population doesn't make a population worse, other things being equal.'

the present. We strangely have no positive welfarist reason under neutrality to pursue these win-win options.[14]

The way neutrality can swallow up gains in welfare is pertinent to Parfit's (1984, 453) pointed question (and yet is easily overlooked).[15]

Compare three outcomes:

(1) Peace.

(2) A nuclear war that kills 99% of the world's existing population.

(3) A nuclear war that kills 100%.

Outcome (2) would be worse than (1), and (3) would be worse than (2). Which is the greater of these two differences?

Parfit himself suggests that most people believe the greater difference is between (1) and (2). The neutral approach to welfare, unlike the totalist approach, can partially respect that intuition. By neutrality, (2) and (3) are incommensurable, assuming the two are identical except that the former includes extra people with worthwhile lives. So (3) is not worse than (2). But at least it is not the case that the gap in welfarist good between (2) and (3) is much greater than the gap between (1) and (2). (Let us assume that the populations in (1) and (2) end up being the same size.)

However, that is not the whole story. The problem with neutrality is that it does not preclude a highly counter-intuitive assessment of options (1) and (3). Most people believe that (3) is much worse than (1), since (2) is much worse than (1). But neutrality allows that (3) and (1) are incommensurable. Consider the shift from (3) to (1): it involves large welfare gains for the small population who were bound to live—from the destitution and shortening of lives that precedes extinction to a state of world peace and increased longevity. And then there are all these extra people living in the state of world peace, since humanity does not go extinct. Surely (1) is better than (3). The trouble is that, on neutrality, the additional persons on the world peace outcome may 'swallow up' the large gains in welfare to the original people. After all, if the destitute people were joined by additional people who enjoyed not just world peace but super blissful world peace, this larger population outcome would have been incommensurable with the smaller destitute population. And plausibly, the outcome (1) is no better a population than the one with super

[14] This is presumably why Broome (2005, 410) claims that neutrality yields 'incredible' assessments of everyday choices: specifically, that doing nothing about climate change, or more locally, about road safety or taxation reform, is no worse than acting in a positive way to mitigate climate change, improve roads or reform the taxation system respectively.

[15] MacAskill (2022, 168) quotes this passage, and Setiya (2022) moreover appeals to Parfit's three options in his criticism of MacAskill's claims.

blissed-out extra people: the downward shift in welfare of many from super blissful world peace to world peace outweighs the upward shift in welfare of relatively few from destitution to world peace.

5. Concluding Remarks

We see that the neutral approach to welfare is more complicated than first impressions suggest. The greediness of neutrality, when there are many additional lives are at stake, means there is less of a contrast between the neutral and totalist approaches to welfare than might be hoped. On either approach, significant present sacrifices in welfare may be permissible (albeit, for neutrality, not required) if the result is a small reduction in the probability of some future threat, including threats whose severity turns on premature human extinction.

If Table 2 more or less accurately describes the choice problem we now face—whether we should aim to directly reduce the risk of some threat—neither the neutral nor the total approaches clearly furnish us with strong reasons *not* to act on the future threat and instead focus on more proximate suffering. The extreme gains in welfare for the 'far-sighted' option under State 1 are relevant (and if large enough, swamp the comparison of options) for both kinds of approach.

One might yet hope that there is *some* plausible neutral approach to welfare that allows us to effectively sideline, in our welfarist reasoning, changes to the timing of human extinction or to population size more generally. After all, the analysis in Section 4 depends on a specific version of the neutral approach. But it is far from assured that alternative versions of neutrality would change the overall story in ways that defenders of neutrality would welcome. For instance, the extent to which neutrality swallows up gains in welfare for an original population could be mitigated by lowering the upper bound of the neutral range to something not too much greater than zero.[16] But this would only lessen the difference between the neutral and totalist approaches to welfare. One might otherwise make different assumptions when it comes to comparing populations of fixed size. Instead of appealing to average welfare, one might rather compare fixed-size populations in terms of minimum welfare, for instance, or in some other way more sensitive to (in)equality. But many such differences will lead to changes in degree rather than kind when it comes to greedy neutrality. Similarly for the assumption of transitivity, I speculate. At any rate, the burden falls on defenders of neutrality to provide a fully worked out account that has clear advantages over the one presented in Section 4.

[16] This would be what Rabinowicz (2009) calls a 'moderate' interpretation of the neutral range. His own critical-range utiltarianism involves a moderate neutral range—it is the range of critical values.

That brings me back to the underlying aim of the paper, which is to deflate the idea that the fundamental approach to welfare one adopts matters a great deal in assessing longtermists' conclusions about future threats that affect the size of the population. The robustness of the longtermists' claims about what we have strong welfarist reason to do and not do hangs on whether the extreme empirical picture that is painted is in fact true. Do our choice problems really look like those described in Tables 1 and 2? If not, then we may *not* have strong welfarist reason to try to directly reduce the probability of premature human extinction, even on a simple totalist account of the welfarist good.

Table 3: Modified Choice Problem for a Varying-Sized Population

	State 1		State 2		…	State n	
	necessary	**possible**	**necessary**	**possible**	**…**	**necessary**	**possible**
short-sighted	modest	none	modest	none	…	modest	none
medium-sighted	high	small	high	small	…	high	small
far-sighted	none	extreme	none	none	…	None	none

What would an alternative choice scenario look like? Table 3 offers just one example; it includes a 'medium-sighted' option. The 'medium- sighted' option might, say, involve institution building that is very good, over the long run, for the given number of people bound to exist, and also enhances expected population size to some extent relative to the status quo. Moreover, the 'far-sighted' option might possibly backfire, making a long-enduring tyrannical regime more likely for the fixed population (an extremely bad outcome). In that case, even on a simple totalist account of the welfarist good, it is not obvious that there is welfarist reason to pursue the 'far-sighted' option.

I have said nothing here to help settle the question of whether Table 3 or Table 2 more accurately represents the choice problem we now face regarding future threats that affect population size. The point is rather that these are the kinds of hypotheses that we should be raising and scrutinising in assessing longtermists' choice conclusions about what we have welfarist reason to do. The totalist approach to welfare plays a relatively minor role in the longtermists' mathematics.[17]

[17] Many thanks to Christian Barry, Hilary Greaves and Alan Hájek for valuable comments on draft versions of this paper. Thanks too to the seminar audience of the Institute for Futures Studies (IFFS) for very helpful feedback. I am moreover grateful for support from the 'Climate Ethics and Future Generations' project at the IFFS, funded by Riksbankens Jubileumsfond (grant number M17-0372:1) and from an ANU Futures Scheme grant.

References

Arrhenius, G. *Population Ethics: The Challenge of Future Generations*, unpublished manuscript.

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

Broome, J. 2005. 'Should We Value Population?', *The Journal of Political Philosophy*, 13(4): 399–413.

Frick, J. 2017. 'On the survival of humanity', *Canadian Journal of Philosophy*, 47(2–3): 344–367.

Frick, J. 2020. 'Conditional Reasons and the Procreation Asymmetry', *Philosophical Perspectives*, 34: 53–87.

Frick, J. 2022. 'Context Dependent Betterness and the Mere Addition Paradox', *Ethics and Existence: The Legacy of Derek Parfit* (eds. McMahan, J., Campbell, T., Goodrich, J., and K. Ramakrishnan) Oxford: Oxford University Press.

Greaves, H. and MacAskill, W. 2021. 'The case for strong longtermism', *GPI Working Paper* No. 5-2021.

MacAskill, W. 2022. *What We Owe the Future*, New York: Basic Books.

Narveson, J. 1973. 'Moral problems of population', *The Monist*, 57: 62–66.

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

Nebel, Jacob M. 2019. 'Asymmetries in the Value of Existence', *Philosophical Perspectives*, 33: 126–145.

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

Rabinowicz, W. 2022. 'Getting Personal: The Intuition of Neutrality Reinterpreted', *The Oxford Handbook of Population Ethics* (eds. Arrhenius, G., Bykvist, K., Campbell, T., and E. Finneron-Burns) Oxford: Oxford University Press.

Setiya, K. 2022. 'The New Moral Mathematics', *Boston Review*, August 15th 2022: https://bostonreview.net/articles/the-new-moral-mathematics/

Thomas, T. forthcoming. 'The Asymmetry, Uncertainty and the Long Term', *Philosophy and Phenomenological Research*.

Voorhoeve, A. 2014. 'How Should We Aggregate Competing Claims?', *Ethics*, 125(1): 64–87.

Melinda A. Roberts[1]

Population, Existence and Incommensurability[2]

In this paper, I consider what highly plausible principles we can consistently combine with certain deeply held, widely shared intuitions we have regarding matters of existence. Does an approach that has such intuitions at its core—does the *existential approach*—rule out, e.g., *transitivity*? Does it rule out *trichotomy*? Is it like certain *incommensurability theories* in those respects? Another question I'll consider is whether our only means of *both* securing deeply held, widely shared intuitions regarding matters of existence *and* avoiding inconsistency lies in incommensurability. As part of those discussions, I'll note what the existential approach has to say about two nonidentity cases, briefly sketch Broome's inconsistency objection against the so-called *neutrality intuition* and finally explore Rabinowicz's incommensurability proposal for avoiding Broome's objection *without* abandoning intuition.

[1] Department of Philosophy, Religion and Classical Studies, The College of New Jersey, Ewing NJ U.S.A 08528; robertsm@tcnj.edu

[2] For their comments on a version of this paper I presented at the Incommensurability and Population-Level Bioethics Conference organized by the Rutgers Center for Population-Level Bioethics and the Institute for Futures Studies (May 2022), I am very grateful to Chrisoula Andreou, Krister Bykvist, Wlodek Rabinowicz and several other members of the audience at that conference. For their important comments and recommendations, both substantive and stylistic, I am very grateful as well to Olle Torpman and Timothy Campbell.

1. Introduction

What I'll call the *existential approach* includes—and gets its name from—the *existence condition* (EC):

> Existence condition (EC): Where x and y are possible worlds and y is accessible relative to x, x is morally worse than y, and a choice c made at x is morally wrong, only if
>
> there is a person p and an alternate accessible world z such that:
>
> (i) p does or will exist in x, and
>
> (ii) x is worse for p than z (where z may, but need not, be identical to y).

EC is a non-additive but still a consequentialist principle—and a maximizing principle at that: if, for a given person who does or will exist at a given world, that person's wellbeing *isn't* maximized at that world as compared against all other accessible worlds, then EC leaves the door open for the result that that world *is* worse than whatever other world we are comparing that world against: even one where the person never exists at all.

The main work of EC is to capture a familiar intuition about existence: that, other things equal, leaving a person out of existence altogether doesn't make things morally worse; the intuition that, if you *want* to make things *morally worse*, you must make things *worse for* a person who does or will exist.

One question I'll address in this paper is what highly plausible principles we can consistently combine with EC to define the existential approach. Can we accept *transitivity*? *Trichotomy*? Or does the existential approach, like certain forms of the *incommensurability approach*, rule out those principles?

A second question I'll try to answer is whether an approach that *both* secures certain highly intuitive results in cases where a person's coming into existence is at stake *and* avoids inconsistency *requires* incommensurability in some form or another.

2. Two nonidentity cases and the existence condition

The first case we shall look at—the *two option nonidentity case*—isn't, I think, a very interesting type of *nonidentity* case. It doesn't, that is, seriously challenge EC.

It's a case in which no additional wellbeing at all is available to be assigned to either of two nonidentical future people. The better off person A *cannot* accessibly be made any better off than *that* person is in w1, and the less well off person B *cannot*

accessibly be made any better off than that person is in w2. Each individual's well-being has been maximized at the world at which that individual exists.

Figure 1: Two option nonidentity

(No accessible world better for B than w2; w1 exhausts the set of worlds accessible relative to w2; c1 and c2 exhaust the set of available choices; connection between choice and world is certain; A and B are distinct persons.)		
	c1	c2
	w1	w2
90	A	
. . .		
50		B
. . .		
0	*B never exists*	*A never exists*

Now, *of course* there exist more remote, *logically* possible, worlds that *are* better for B. The accessibility stipulation just reflects the fact that it's *part of the case* that those worlds can't come about *given* certain features of w1 and w2: the history of w1 and w2 (and the fact that we can't change the past); given gravity (and the fact that we can't undo gravity); given genetics (and the fact that we can't "correct" certain genetic or chromosomal features without destroying the individual). It's not just that such a better-for-B world is *highly improbable* or that such a world (counterfactually) *would* not have come about "but for" whatever agents in fact did to bring about w2. Rather, it's that such a world *isn't available*—isn't, that is, *accessible*: it isn't a world that agents, whether operating either as individuals or collectively, have the ability, the power, the resources to bring about.[3]

Applied to this case, EC implies that it's not the case that w1 is worse than w2 (~w1 $\prec$ w2) and it's not the case that w2 is worse than w1 (~w2 $\prec$ w1).

Let's now add trichotomy to the picture.

[3]Accessibility isn't otherwise defined for purposes here. When our focus is moral betterness, however, it seems plausible that a world y is accessible relative to a world x just in case agents, regardless of the probabilities and regardless of what some or all of them would have done had they not done what they in fact have done in x, have the ability, the power, the resources to bring about y rather than x. In contrast, when y itself is ruled out by features of x that are themselves unalterable relative to x (x's past; the fact of gravity), then y is, though barring inconsistency logically possible, nonetheless inaccessible. Why bring agents into the picture? Because the concern here isn't aesthetic value; or prudential value. It's moral value: the logically possible world that allows us all to live forever at our highest possible wellbeing levels throughout doesn't rank, in respect of its moral betterness, any higher than the poor and morally bankrupt world in which we do now find ourselves.

> *Trichotomy*: If it's not the case that x is worse than y (~x ≺ y) and it's not the case that y is worse than x (~y ≺ x), then x is exactly as good as y (x ~ y).

EC, trichotomy and some conceptual principles now together instruct that w1 is exactly as good as w2 (w1 ~ w2).

I find that result at least credible—and certainly no counterexample against EC. After all, in *all* the *interesting* nonidentity cases—the cases that, on their face, challenge the intuition behind EC (*though not I think successfully due to fallacies in the relevant probability assessments*[4]); cases like depletion, risky policy, historical injustice, climate change, slave child and pleasure pill—things *could* accessibly have been made better for the less well off person. But, in two option nonidentity, things *could not* accessibly have been made better for that person; and it's just part of the case that whether w1 or w2 obtains makes no difference (in terms of either existence or wellbeing) to anyone else at all. It's hard, then, for me to see that w2 is *morally worse* than w1, that is, that w1 is *morally better* than w2.

I reject, in other words, what we might call *procreative perfectionism*.

In the absence of further accessible alternatives, and other things equal, leaving a person, even the better off person A, out of existence and bringing the less well of person B into existence, doesn't make things worse.

Moreover, if we accept that, *under conditions of certainty*, the evaluation of choices as permissible or wrong is closely connected to the ranking of worlds in respect of their moral betterness—if we accept the *principle of connection*—we'll conclude further that c1, which ends in w1, and c2, which ends in w2, are both permissible. (That particular implication from connection is actually built into EC.)

But now consider *three option nonidentity*, a more interesting nonidentity case—one that in one way looks a little more like depletion, risky policy, historical injustice, climate change, slave child and pleasure—with one huge, morally critical, difference: *in contrast to all those cases, three option nonidentity stipulates that the relation between choice and world is certain: that, given a choice c, the probability that the identified world w will unfold is 1.0.*

[4]Roberts 2007; Roberts 2009; Roberts and Wasserman 2017.

Figure 2: Three option nonidentity

(The world w3 is accessible and better for B than w2; A and B are distinct persons; w1 and w3 exhaust the set of worlds accessible relative to w2; c1, c2, c3 exhaust the set of available choices; connection between choice and world is certain.)			
	c1	c2	c3
	w1	w2	w3
90	A		
. . .			
60			B
50		B	
. . .			
0	*B never exists*	*A never exists*	*A never exists*

Here, B's future may unfold, if B exists at all, in *more* than one way: a way that is *worse* for B and a way that is *better* for B. In two option nonidentity, EC implies that w1 ~ *w2*. EC now tells us that w1 ~ *w3*. But EC—unlike many other so-called "person-affecting" principles—*doesn't* imply, in three option nonidentity, that w1 ~ w2. The inference to that result is blocked: for, in the second case, there is a w3 that is better for B than w2, and we thus can't infer from EC that w2 *isn't* worse than w1.[5]

We therefore have room to say that w2 *is* worse than w1 (w2 ≺ w1).

And we proceed to do just that. We appeal to the substantive moral principle we can call *same people Pareto* as well as trichotomy and some other conceptual principles.

> *Same people Pareto*: If exactly the same people do or will exist in worlds x and y, and y is better for at least one of those people than x and worse for none, then x is morally worse than y (x ≺ y) (that is, y is morally better than x (y ≻ x)).

About the second case we can then say:

1. w2 ≺ w3 same people Pareto
2. w1 ~ w3 EC, trichotomy
3. w3 ⪯ w1 logic, 2
4. w2 ≺ w1 transitivity, 1, 3

[5] We can spell that out. According to EC, w2 ≺ w1 only if there does or will exist a person in w2 such that w2 is worse for that person than some other accessible world. In the second nonidentity case that condition is satisfied; w2 is worse for B than w3; the inference from EC that w2 isn't worse than w1 is therefore blocked; and we room to say that w2 is worse than w1—that w2 ≺ w1.

I find the result that w2 is worse than w1 highly intuitive. (Yes, it means rejecting the *mere addition principle*. But that principle has become increasingly suspect in recent years.)

Connection lends that account of the case intuitive force. It implies that c1 and c3 are permissible and that c2 is wrong. You can bring a child into existence or not, but if you do bring a child into existence and can do (or could have done at some point in the past) more for that child rather than less, then you ought to do (or ought to have done) more for that child rather than less. One *need* not conceive a child, or continue the early or middle pregnancy that will end in a new person coming into existence. (Here, I make the plausible assumption that there's no *person* there until a *connected consciousness* has materialized, which seems not to happen until at least the 24th week of pregnancy or so.[6]) But if one *does* bring a new person into existence—say, one's own child—then one *must* (other things equal) do more for that child rather than less: stockpile resources months or years or decades before conceiving a child (assuming doing so contributes to the future child's wellbeing); take vitamins before conceiving a child (assuming doing so contributes to the future child's wellbeing); avoid certain drinks and drugs early in the pregnancy (assuming doing so contrbutes to the future child's wellbeing).

Two questions about the existential accounts of our two nonidentity cases arise.

Consistency. First, don't we now have an inconsistency—one that forces us to reconsider EC itself or perhaps trichotomy or transitivity or something else near and dear to our hearts? Don't our two nonidentity cases nicely demonstrate that the *existential approach is inconsistent*?

Not, I think, if we understand that what we are comparing in these two cases or any of the cases under scrutiny in this paper *aren't* simple distributions of wellbeing across particular populations but rather *worlds*.[7]

Since worlds have all their features necessarily and since features like the *accessibility of a world z relative to a world x and a world y* are built into x and y, the w1 and w2 in *two* option nonidentity are *necessarily distinct* from the w1 and w2 in *three* option nonidentity. Just because worlds *happen* to share a wellbeing distribution across a given population *isn't* enough to make them *the same world*.

[6] Roberts 2010.

[7] Some theorists have assumed that the only morally critical feature about a given world is its population and the distribution of wellbeing across its population. The existential approach denies that minimalist picture. While it avoids any appeal to intention, character, special prerogatives, reasons, duties, rights and the like, it nonetheless insists that what is going on in worlds accessible to a given world x—with accessibility itself being a function, as noted just below, of what is going on in x itself—on occasion bears on the moral value of x. Specifically, it may on occasion be relevant to how x compares against y that a further accessible world z makes things better for a person who does or will exist in x.

To make this point explicit, we can adopt a more precise vocabulary to describe the two cases. We can, for example, just add little asterisks to one pair of "w1" and "w2" or the other. Thus for two option nonidentity we can write w1* ~ w2*, and for three option nonidentity we can write what we did before: w1 ~ w3, and w2 ≺ w1 and w2 ≺ w3.

Inconsistency avoided—and same people Pareto, transitivity and trichotomy preserved—we can move on to a second question.

Calibration of strength of necessary condition by reference to accessibility. What justifies—what's the reason for—calibrating EC's necessary condition on moral worseness in terms of *accessibility*? Why do we say that x can be worse than y *only if* an *accessible* world *z* makes things better for a person who does or will exist in x?

The reason EC includes that condition rather than any stronger or a weaker condition is that those alternate conditions don't work within a framework that aims to capture the relevant intuition about existence: that aims to say that leaving a person out of existence altogether doesn't, other things equal, make things worse.

The more typical "person-affecting" way of doing things would include a *stronger* condition. It would have it that x is worse than y *only if* x is worse for a person who does or will exist in x *than y is for that same person*. But the forced disregard of z *can camouflage* moral deficiencies in x and that make x worse than y: it can camouflage facts about x that bear on the moral value of x. When existence is at stake, it's often a world beyond x and y that highlights for us what is *amiss* about x and what makes x *worse* than y.

We see that in three option nonidentity: it's w3, not w1, that reveals the moral deficiency in w2; it's what is going on in w3 that correctly, I think, blocks the inference to the unhappy result that w2 *isn't* worse than w1 and thus opens the door to the very happy result that w2 *is* worse than w1.

Moreover, amending EC to include the more typical, and stronger, condition leads to the sort of inconsistency we *can't* avoid by simply adopting a more precise vocabulary. Amending EC to include the stronger condition would tell us, not just that w3 is exactly as good as w1, but also that w2 is exactly as good as w1. The theory that also adopts same people Pareto, transitivity and trichotomy really is, then, enmeshed in inconsistency. We avoid the inconsistency by avoiding the result that w1 ~ w2—and say instead what we said before: that w2 is *worse* than w1.

What of the alternative of amending EC to include a *weaker* condition? That doesn't work either, I think, for two reasons. First, such an approach would have it that x is worse than y *only if* x is worse for a person who does or will exist in x than *any* world z in that vast collection of worlds that is *logically possible* relative to x whether accessible or not—worlds whose histories are quite unlike x's; worlds where gravity fails; worlds where happy people live on forever. *Any* such world z would

reveal a damning *moral* deficiency in x and thus open the door to the result that x is worse than y.

To see that that's so, we can just go back to three option nonidentity. The condition on w3's being worse than w1 is immediately satisfied by the fact of all of those many *inaccessible* but *logically possible* worlds—w4, w5, w6, w-sub-a-hundred-zillion—that make things better for B than w3. If the progression of worlds that are better for B than w3 extends indefinitely but ends finally at some fixed point w_n, then the inference under amended EC to the result that *w3, w4, w5 and all those worlds short of* w_n isn't worse than w1 will be blocked, And same people Pareto, transitivity, trichotomy and logic will step to say that *all* those worlds, including w3, are *worse* than w1. That seems extreme (even to me).

And there's a second problem as well. It seems clearly permissible (though not obligatory) to bring B into existence in w3. After all, it's part of the case that w3 represents the very best that agents can accessibly do for B. Now, it's true that we can secure that nice permissibility result by modifying connection: we can move accessibility away from the job of world-ranking back to its more traditional job of choice-evaluation and we can then say that choices that end in worlds than which there is no better *accessible* world are *all* permissible. But that way of doing things creates a new problem: that, even though w3 is morally *worse* than w1, the choice that ends in w3 is perfectly *permissible*. It's one thing to say that choices that end in some logically possible better world aren't *obligatory* if they are inaccessible. But in three option nonidentity w1 *is* perfectly accessible *and* is *better* than w3. How, then, can the choice that ends in the perfectly accessible and morally better w1 *not* be obligatory? How can the choice that ends in w3 *not* be wrong?

Thus the *existential* approach. We now turn to the *incommensurability* approach.

3. Incommensurability approach

3.1. Broome's case

Three option nonidentity closely tracks the case John Broome used to show that what he calls the *neutrality intuition* is false.[8] (It's that intuition that Wlodek Rabinowicz—as we shall see—aimed to rescue from inconsistency through his theory of incommensurability.)

First, the neutrality intuition itself—the formulation of the underlying intuition

[8] See Broome 2004, pp. 145–149. Broome's primary interest is in comparing worlds in respect of their overall betterness. Choices, however, have been included in Figure 3 below ("Broome's Case") since Rabinowicz's treatment of the case, as well as my own, extends to the evaluation of choice.

(that is, the so called "person-affecting" intuition) that Broome himself presents as the most charitable he can come up with. (It's not that charitable.)

For a certain range of wellbeing levels—not just a *single* wellbeing level, but rather a *range* of wellbeing levels; the *neutral range*—the neutrality intuition states that:

> *Neutrality intuition*. If an additional person's existence in a world y makes y neither better nor worse *for anyone else* than a world x *and* x and y otherwise contain the same people, then that person's existence in y at a wellbeing level in the neutral range doesn't make y *morally better* or *morally worse* than x.

The case that Broome relies on in his argument to the conclusion that the neutrality intuition is inconsistent is this (the fiction of *Harry* is my own):

Figure 3: Broome's case

(w5 and w6 exhaust the set of worlds accessible relative to w4; c4, c5 and c6 exhaust the set of available choices; connection between choice and world is certain.)			
	c4	c5	c6
	w4	w5	w6
+60			Harry exists
+50		Harry exists	
+0	*Harry never exists*		

According to the neutrality intuition, w5 isn't either better or worse than w4, nor is w6. On those facts, trichotomy tells us that w5 is exactly as good as w4, and so is w6. Still other principles—transitivity and symmetry—instruct that w6 is exactly as good as w5. Finally, same people Pareto produces the inconsistency: surely w5 is worse than w6.

Amongst all these principles, Broome considered the neutrality intuition itself to be the weak link. And he rejects it, putting in its place the claim that there exists, not a *range* of neutral wellbeing levels, but rather *exactly one* neutral wellbeing level: exactly one level such that, other things equal, an additional person's existence at that level makes the one world neither better nor worse than the other. At *every other level*, the additional person's existence makes things either better or worse.

* * *

It can't be emphasized enough how deeply at odds with intuition Broome's conclusion that there's just a single neutral level of existence is.

Suppose that the single neutral wellbeing level is what Harry has in *w5*. Then, w6 is morally better than w4. Connection implies that c4, the choice *not* to bring Harry into existence, is wrong, and c6 *obligatory*. *My* intuition that c4 is just fine—that it's *perfectly* fine to leave Harry out of existence—remains alive and well.

Or suppose that the single neutral level is what Harry has in *w6*. Then—in another case, a case just like Broome's original *except* that *there is no accessible w6*; there is just, we can say, w4* and w5*, such that Harry now has more wellbeing in w5* than in any other accessible world—w5* is *worse* than w4*. Given connection, we conclude that c5 is wrong. Here's another intuition: that's just false. When agents—all of them, as individuals or collectively—have done the best they accessibly can for a child whose existence is worth having, they haven't done anything wrong.

Broome's single neutral level proposal avoids the inconsistency. That's a plus. But to *solve the problem* that Broome's case gives rise to—that the existence gives rise to—requires, I think, *more* than avoiding an inconsistency. To solve the problem is to solve the *puzzle*: it's to fit *all* the puzzle pieces together and not just fit *some* of them together while tossing the rest. Of course, even a deeply held, widely shared intuition is on occasion quite rightly rejected or at least amended. But to solve the puzzle is to come to understand just how our original intuition has gone wrong. That, in turn, requires a *new platform* of deeply held, widely share intuition, a platform that serves to loosen the hold that the original intuition had on us. I don't think Broome's analysis provides us with the requisite platform.

4. Rabinowicz's incommensurability proposal

4.1. How incommensurability avoids Broome's inconsistency argument

To get the inconsistency, we need the neutrality intuition and trichotomy.

Where Broome considered the neutrality intuition the weak link, Rabinowicz proposes that it's trichotomy. If we reject trichotomy in favor of *incommensurability*, we can provide an interpretation of the neutrality intuition that avoids inconsistency but doesn't force us—*try* to force us—to reject *any* of the various parts of the neutrality intuition.[9]

[9] Rabinowicz 2009.

> *Core incommensurability principle*: x is *incommensurate* with y iff
> it's not the case that x is worse than y (~x ≺ y) *and*
> it's not the case that x is better than y (~x ≻ y) (i.e., not the case that y is worse than x) *and*
> it's not the case that x is exactly as good as y (~x ∼ y).

Let's go back to Broome's case. When the addition in question is a *mere* addition—as in Broome's case—Rabinowicz proposes that the world "with added people at neutral levels must be incommensurate with the world without these additions."[10]

> *Additional person incommensurability principle*. When the addition in question is a *mere* addition—in any case where the addition doesn't make things better or worse for anyone else—the world "with added people at neutral levels must be incommensurate with the world without these additions."

On this principle, w5 is incommensurate with w4, and so is w6. And—since incommensurability isn't considered transitive—we never get to the problem result that w5 is incommensurate with w6. We can thus consistently retain same people Pareto and take the position that w5 is morally worse than w6.

Stirring connection into the mix only strengthens the claim that the incommensurability approach comports with intuition. Since w5 ≺ w6, we conclude that c5 is wrong; and since there exists no world that is better than either w6 or w4, we can also conclude that c6 and c4 are permissible.

I, at least, find *most*—though not *all*; I'll come back to that point shortly—of those results highly intuitive. Rabinowicz, in other words, as compared against Broome comes much closer to actually *solving* the puzzle rather than *ignoring* the puzzle—rather than throwing some of the puzzle pieces into the fire.

4.2 The existential account

The existential account of Broome's case is so like the existential account of three option nonidentity that we don't need to work through its details here.

To sum it up: w4 is exactly as good as w6, while w5 is worse than both; c4 and c6 are permissible and c5 is wrong. We as before are under no pressure to reject transitivity or trichotomy, and end, as far as I can tell, with a perfectly consistent account of Broome's case.

[10] Rabinowicz 2009, p. 392.

4.3 Three distinctions

If the permissibility results from the existential and the incommensurability approaches are the same, on what grounds do we prefer one rather than the other?

4.3.1 Distinction as to whether w5 is worse than w4

Though they agree on their evaluations of the relevant choices, the incommensurability approach and the existential approach part ways on their rankings of the relevant worlds.

The incommensurability approach stays true to the neutrality intuition. It implies that w5 isn't worse than w4—and, more generally, accepts *all* parts of the neutrality intuition.

In contrast, the existential approach instructs that w5 is worse than w4, which is to reject *one* part of the neutrality intuition while accepting the rest.

So: is w5 worse than w4, or not? Perhaps we can go either way on this point.

4.3.2. Distinction as to trichotomy, transitivity

The existential approach preserves trichotomy and has no problem with transitivity. While we can accept that there are reasons to think trichotomy fails in many contexts (when, e.g., different values are at stake), it's less clear that it fails in the contexts we've considered here. I take it that significant controversy swirls around the rejection of trichotomy only if the claim is made in respect of a single value, e.g., moral value, and not a plurality of values. Surely, after all, we find claims of incommensurability at least plausible (though we arguably also have the option of saying that we just aren't yet in possession of a complete theory) when we are comparing baskets of values of very different sorts: filial duties against duties of citizenship; aesthetic value against prudential value; prudential value against moral value. Once, however, we narrow our focus to questions involving a single value—moral value—and ask the single question of how two worlds compare in respect of moral value (i.e., whether one world is morally worse than another), it becomes conceptually hard to see how trichotomy (or transitivity) might fail. If x isn't better than or worse than y, how can x not be exactly as good as y? Where else is there to go?

4.2.3. Distinction as to the difference proposition

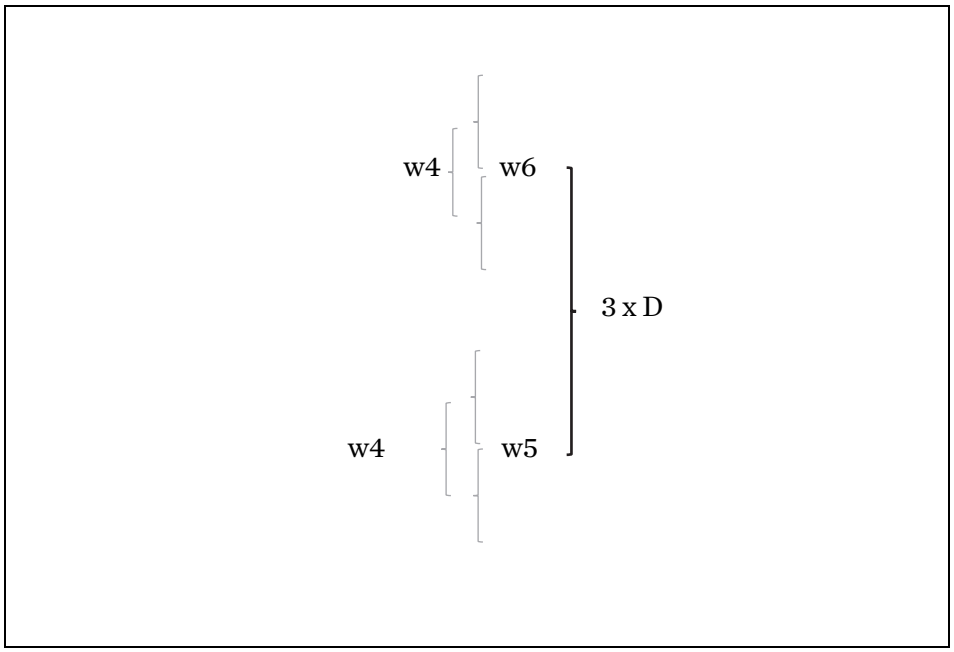
The following proposition seems plausible:

> *Difference proposition*. If the incommensurability approach is correct, then there nonetheless exists some *difference* (or, graphically, some *distance*) D between values such that, for any two worlds x and y, if the difference (graphically, distance) in value between x and y is *greater* than D, then either x is worse than y or y is worse than x.

Let's look at how this proposition works in the context of a *certain version of* Broome's case, a case that includes certain details that our original presentation of the case lacked. Let's suppose that the difference, or distance, in value between w5 and w6 is at least three times D. This estimate reflects the fact that, despite the incommensurability rife in the case, w5 and w6 remain fully commensurate. Of course, *any* improvement in Harry's position from w5 to implies, under the incommensurability approach, that w5 is worse. But I take it that even a minute improvement in Harry's position is perfectly consistent with the position that the distance in value is not minute at all but rather substantial: it's enough to ensure that w6 is indeed *better* than w5 and not *incommensurate* with w5. At the same time, surely the distance in value between w5 and w6 increases as things are made still better for Harry. I thus take it that improving Harry's position by a full ten units—by one-fifth—makes it safe to say that the difference in value between w5 and w6 is at least three times D. (If that guess doesn't seem apropos, we can *adjust* certain details of the case: we can increase Harry's wellbeing in w6 until the point at which the distance in value between w5 and w6 *is* at least three times D.)

The difference proposition implies further that, since w5 and w6 are both incommensurate with w4, the distance between w5 and w4 is *equal to or less* than D *and* the distance between w6 and w4 is *equal to or less* than D. Graphically: w4 hovers in value around w5, while w4 also hovers in value around w6.

Figure 4: Geometry of incommensurability



Fine. The problem is that, under the geometry we've laid out, the distance between w4, hovering around w5, and w4, hovering around w6, is of necessity greater than D—thus that w4 is either better than or worse than w4. But w4 is, *surely*, exactly as good as w4. And that's an inconsistency.

Does this mean that the incommensurability approach is inconsistent? No. It just means that that approach can't be understood by reference to the difference proposition. But we then seem left with the following picture: w6 is clearly *better* than w5 but it's nonetheless *no farther apart* in value from w5 *than w4 is*: w5 and w6 might be no more *different*, or *distanced*, from each other in terms of moral value either is from w4.

I think that that's an implication that's hard to grasp. At the very least, it means that we can't understanding incommensurability without replacing the more graphic, more mathematical, concepts and principles we are accustomed to employing in evaluating worlds with some other metaphor altogether. Will that metaphor fully enlighten us as to what is going on in the many additional person cases we need to understand—not just Broome's case, but also our two nonidentity cases and still other cases as well? I'm not sure that it will. Even if it doesn't, the implication might be that incommensurability is a hard theory to *grasp*, not that it's *false*. But perhaps we already knew that it is a hard theory to grasp.

5. Conclusion

I started this paper with the question what highly plausible principles we can consistently combine with EC. Does an approach that has EC at its core rule out *transitivity*? Does it rule out *trichotomy*? Is the existential approach like the *incommensurability approach* in those respects? I've argued that EC is consistent with transitivity and trichotomy.

My second question was whether an approach that *both* secures certain highly intuitive results in cases where existence is at stake *and* avoids inconsistency *requires* that we accept incommensurability. If the existential approach is itself viable, I've shown that it doesn't.

References

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

Rabinowicz, Wlodek 2009. "Broome and the Intuition of Neutrality." Philosophical Issues 11: 389–411.

Roberts, Melinda A. 2007. "The Nonidentity Fallacy: Harm, Probability and Another Look at Parfit's Depletion Example," Utilitas 19, 267311.

Roberts, Melinda A. 2009. "The Nonidentity Problem and the Two Envelope Problem." In Harming Future Persons, eds. Melinda A. Roberts and David T. Wasserman. Springer. Pp. 201–228.

Roberts, Melinda A. 2010. Abortion and the Moral Significance of Merely Possible Persons: Finding Middle Ground in Hard Cases. Springer.

Roberts, Melinda A. and David T. Wasserman (2017). "Dividing and Conquering the Nonidentity Problem," in Current Controversies in Bioethics, eds. Matthew Liao and Collin O'Neil (Routledge), pp. 81–98.

Patrick Kaczmarek[1] and SJ Beard[2]

Do We Owe the Past a Future? Reply to Finneron-Burns[3]

Many of our forebears went beyond the call of duty, sacrificing much, to benefit those who would come after them. Elsewhere, we claimed that preventing human extinction renders those past sacrifices more worthwhile. We also developed the Unfinished Business Account, which together with the first claim implies that we wrong past people by squandering their sacrifices, even though this failure is in no way worse for past people. Elizabeth Finneron-Burns has recently questioned both of these claims. This note responds to her criticisms.

[1] Centre for the Study of Existential Risk, Cambridge, pk496@cam.ac.uk

[2] Centre for the Study of Existential Risk, Cambridge, sjb316.cam.ac.uk

[3] This paper was made possible through the support of a grant from Templeton World Charity Foundation, Inc. TWCF0367, A Science of Global Risk. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of Templeton World Charity Foundation, Inc. We are grateful to Elizabeth Finneron-Burns for her paper and for continuing discussion on this topic. We are also grateful to Dale Miller and an anonymous reviewer for feedback on earlier versions of this paper.

Many of our ancestors went beyond the call of duty, sacrificing much, sometimes even making the ultimate sacrifice, to benefit those who would come after them. They fought for democracy, freed slaves, built the infrastructure that serves us and so forth. Elsewhere, we both claimed that preventing human extinction renders those past sacrifices more worthwhile and developed the Unfinished Business Account, which states that if actor p reasonably judges performing a supererogatory act ϕ at great sacrifice to herself will enable beneficiary q to achieve a greater good, then failure to promote the good made possible by ϕ wrongs p (Kaczmarek and Beard 2020, 202). We showed that these claims together imply a *pro tanto* duty to render the sacrifices of past (and present) people, from which we benefited, more worthwhile by preventing human extinction.

Elizabeth Finneron-Burns questions whether it follows from the Unfinished Business Account that we have a duty to prevent human extinction (Finneron-Burns 2021). She develops two lines of criticism. First, she maintains that one cannot derive obligations from worthwhileness. Second, she argues that we beg the question by assuming that future people would be benefited if caused to exist and have a good life. This note responds to her criticisms.

1. On Trust and Worthwhileness

Our suggestion was that when one person makes a sacrifice for another person's good, they entrust that person with a duty to get as much value as possible from their sacrifice (Kaczmarek and Beard 2020, §2). Finneron-Burns raised two arguments against this claim, which we address in turn.

The first is that there may not even be *pro tanto* wrongness involved when one person, the beneficiary, does not do everything they could to maximise the benefits that were made possible by another person's sacrifice. Specifically, Finneron-Burns denies that acting in this way violates any kind of trust since trust requires prior agreement between the parties involved (Finneron-Burns 2021, 6). Because no such agreement is possible between us and our forebears, it cannot be the case that we betray their trust by squandering the benefits made possible by their sacrifice. Finneron-Burns likens this to the case where I gift a million dollars to my neighbours in the hope they set up a college fund for their children. She claims that my neighbours do not betray my trust if they instead used this money to revamp their kitchen.

We interpret her case somewhat differently. My neighbours do betray my trust but they do so *with adequate justification*. I trusted them in the sense that, when I parted with my fortune, I expected that my neighbours would try to do what they ought to do, which was to make the most of my sacrifice. This sort of trust doesn't

depend on prior agreement, and the wrong-making property is not acting contrary to my wishes but instead disregarding the reasons that spring from my sacrifice. However, it is also true that, by forgoing such renovations for the benefit of their children's education, they would be making a sacrifice of their own, which might be too much for me to ask of them. Although they possess reason to set up the college fund, perhaps my neighbours also desperately crave a change of lifestyle, and so exercise their agent-centred prerogative to make their dream kitchen.

In general, when I leave "unfinished business" for others, I do so expecting them to be sensitive to moral reasons, and not to my wishes, dreams, and hopes. One could, of course, question whether 'trust' is the right term for describing this sort of attitude taken towards duty-bearers. However, we take it that our main thought experiment, Liver Transplant, adequately demonstrated that common sense recognizes that certain sacrifices are reason-giving.[4]

This brings us to the second issue raised by Finneron-Burns. She maintains that, even if the source of the wrongness is a betrayal of trust, trust comes with limits. We recognized a number of such limits in the original paper; for instance, writing that "it is clearly not possible to oblige another person to accept a greater sacrifice than that which one originally accepted" and also that "if our obligations to the past are to play a decisive role in our moral choices, then they should be at least broadly in line with our long-term interests and consistent with our conception of the good life" (ibid, 202). As we understand them, these limits relate primarily to the size and nature of the costs that could be justifiably imposed on the duty-bearer in rendering their benefactor's sacrifice more worthwhile.[5] Indeed, it is what motivates our conclusion about Finneron-Burns' case involving the million-dollar gift to one's neighbours.

Finneron-Burns, on the other hand, seems only to be interested in limits on our obligations based around whether we can render some sacrifice worthwhile or not, and thus wrap up the "unfinished business" handed down by the past. Her argument assumes the need for a clear cut off between sacrifices that are 'worthwhile' and 'not-worthwhile', and she interrogates various such cut-off points based on differing interpretations of what it means to 'realise the full benefits of the sacrifice'.

However, we see no obvious reason to commit to this binary framework. We suspect the disagreement between Finneron-Burns and us on this point stems, at

[4] *Liver Transplant*: Through no fault of his own, Jeff is very sick. He desperately needs a liver transplant. Though he is not obliged to do so, a stranger called Michael gives Jeff part of his liver at the cost of reducing his own lifespan by ten years. After the procedure, Jeff drinks heavily, and he dies from cirrhosis four months later (Kaczmarek and Beard 2020, 200).

[5] F. M. Kamm refers to the costs that duties can justify imposing on the duty-bearer as the 'efforts standard', which she describes being one of (at least) two dimensions of the normative strength of moral reasons alongside the 'precedence standard', which concerns the relative weights of duties when they clash. See (Kamm 1985).

least in part, from a mismatch in how the concept of 'worthwhileness' is being understood. Finneron-Burns writes that "The authors argue that a sacrifice is not morally worthwhile if the beneficiary fails to 'realise the full benefits of the sacrifice'" (Finneron-Burns 2021, 3). But that is not what we think, though we blush to admit this was muddled in the original paper, and Finneron-Burns was right to put pressure on us to clarify. Our view is that a sacrifice would be less morally worthwhile than it might have otherwise been if its full benefits weren't realized by the duty-bearer. The worthwhileness of a given sacrifice is a matter of degree, and it would always be better (in a reason-implying sense) if a sacrifice were rendered more worthwhile, no matter how worthwhile it may have already been made.

One reason someone might believe that worthwhileness is not a matter of degree, but a black and white notion, would be if she believes there is a direct connection between whether we can be obliged to secure benefits from past sacrifices (or, more crudely, whether it would be wrong for us not to realize them) and whether those sacrifices would be made worthwhile by our intervention.[6]

This is not the view that we hold. As rehearsed in the preceding paragraphs, to our minds, beneficiaries have obligations to the past when the reasons that flow from their benefactor's sacrifices are sufficiently strong and decisive, and that this will depend upon at least three things: (1) how worthwhile the sacrifice would be with or without our intervention, (2) the cost to the duty-bearer of intervening and (3) the degree of fit between such an obligation and our disposition to bare such burdens even if they are not obligatory. And so, we simply have no need for a sharp cut-off between worthwhile and not-worthwhile sacrifices but instead obtain unambiguous claims about obligation and supererogation from a non-binary notion of 'worthwhileness' combined with these other considerations.

However, Finneron-Burns' critical discussion did prompt us to consider situations where one or more conditions in the Unfinished Business Account might not hold. We have come to believe that our initial formulation may be too weak; it doesn't apply in cases where the benefactor chooses to give up more than could possibly be gained from that sacrifice. But it seems like perhaps it should. Wouldn't a beneficiary wrong a benefactor (if only non-decisively), even in the case where the upper-bound of value made possible by the benefit that could be achieved was less than the cost of their sacrifice, if the beneficiary still chose to waste that sacrifice for less benefit than it might have realized? Common sense tells us that such suboptimal sacrifices are at least permissible, and even honourable. After all, it's my good, and if I want to

[6] One could deny the very possibility of wrong-making properties featuring in acts that are permissible on balance. This position strikes us as awfully strong, implausible, and we hesitate to ascribe it to Finneron-Burns. Most will agree that an aspect(s) of an action can be *pro tanto* wrong even if the act is a permissible object of choice on balance, as set out in (Chappell 2021, §2.4).

let another person catch a break, then common sense instructs that I should be permitted to do so (Lazar 2019; cf. Hurka and Shubert 2012; Sider 1993). We do not attempt that project here, but it does seem to be a promising place to dig in more.[7]

2. On Begging the Question

In the second half of her paper, Finneron-Burns starts with the following observation. Since in our thought experiment, Liver Transplant, the benefit realized is a benefit to some individual, the benefit that is realized in the case we are concerned with, human extinction, should similarly, by analogy, be understood as a benefit that goes to some individual or thing. She proposes two accounts of who this beneficiary might be: 'humanity', in which case she says that the benefit must be cashed out as extending humanity's lifespan; or the people who might then come into existence with good lives in the future (Parfit 2017, 129).

On the first of these options we submit that Finneron-Burns' claim about this benefit boiling down to longevity is too strong. Johann Frick, for instance, defends a richer notion of humanity's 'final value', which while understood to attach to the species as a whole, and constitute a reason for promoting human survival, is not merely reducible to the longevity of human existence. As he puts it:

> Imagine a world in which each generation of humans dies and vanishes without trace before the next one is born.... Each new generation lives without knowledge of previous generations of humans. The human species survives in this scenario, but a lot of what we mean by 'humanity,' and a lot of what seems uniquely valuable about it — our sense of history, cultural traditions, relationships between parents and children, etc. — is lost (Frick 2017, 362-3).

Similarly, a single human existing for two billion years, might also contribute less to the final value of humanity than six billion people existing together over the next five thousand years, which is the very claim Finneron-Burns offers for rejecting her notion of benefitting 'humanity'.

However, our main point of contention is with the second option, which Finneron-Burns correctly claims is the one that we are more sympathetic towards. She suggests our argument assumes that possible people can be benefited by being caused to exist and have a good life, and that doing so begs the question against our target audience.

[7] A quick-fix would be to replace "a greater" with "some" in our statement of Unfinished Business but there might be other complications.

It is certainly true that we assume that creating people with good lives is good for them. But does that beg the question against those who endorse the No Complainants Claim?

We think not. The overwhelming majority of moral theories aren't expressly opposed to, let alone fundamentally incompatible with, the possibility of benefiting people by doing what is good for them, by causing them to exist and have a good life, even when this is not better for them.[8]

What's more, this is true of moral theories within the No Complainants tradition, which states that an act cannot be wrong unless there is or will be someone whom this act wronged (Parfit 2017, 136). The philosophers crafting these theories have tended to pay little mind to existential benefits because they are concerned about whether anyone has a "complaint" and they believe these sorts of benefits cannot be the source of such complaints, because failing to provide them wrongs nobody. For such moral views, "it is enough to do nothing that would be bad for these people. We could achieve this moral aim in a purely negative way, by doing nothing" (Parfit 2017, 137). But notice that their rather stern (and, to Parfit's mind, impoverished) focus on non-maleficence is consistent with the possibility of existentially benefiting.

Far more controversial is the follow-up claim, that existential benefits give rise to deontic directives. Our target audience, as set out in the paper, are those who accept the No Complainants Claim and thereby resist this further step. It would beg the question against this group to appeal to a moral reason to promote existential benefits. But we did no such thing. Rather, we simply claimed that the failure to produce such benefits could make certain harms from past sacrifices worse by making the sacrifices they were associated with less worthwhile. On the view we put forward, it is solely for the sake of the past people who made these sacrifices (and the complaints they might have against us) that we pursue a future wherein those sacrifices are made most worthwhile. In this way, we were providing an argument for why certain benefits enjoyed by future people may be morally salient even to those who would reject standard arguments in favour of existential benefits being reason-implying.

Our argument could still be said to beg the question against those who vehemently oppose the very possibility of an act being good for someone if it is not also better for her.[9] How worrisome is this for our project?

Not very. While the arguments advanced in support of the stronger view have

[8] For instance, we have elsewhere shown that Scanlon-style Contractualists aren't fundamentally committed to their denial (Beard and Kaczmarek 2019).

[9] That is, those who accept the Narrow Deontic Principle: an act cannot be wrong if it would be worse for no one (Parfit 2017, 119).

some force, we venture that few will end up denying the possibility of existential benefits for two reasons.

Firstly, we standardly make sense of the wrongness of creating a miserable life by appealing to the corresponding notion of 'existential harm'.[10] If we think that this makes sense, as in fact most do, it seems *ad hoc* to insist that existential benefits are downright hokum (Harman 2009, 781–2).

Secondly, the stronger view appears to imply that the vast majority of those now alive have no reason to be grateful for, say, those who worked to prevent a nuclear exchange during the Cold War.[11] After all, had such a war occurred, many of us would not have been born, and thereby we cannot be said to be better off than we would have been had global war not been averted. Yet, we are grateful to these people, and this gratitude does not seem misplaced. The same can be said for the eradication of smallpox, the end of slavery in the Antebellum South and so forth. If not because these things were good for us, what might explain our gratitude?[12]

A more pressing question for our project is whether existential benefits can still give us non-moral reasons for acting. We believe that they can. Understanding that life could be wonderful, even if only because one recognises that it would be wonderful to live such a life, or that future humans could achieve some great goods, even if only because we recognise that these goods would be great, is all that it takes. Each of us can recognise these lives as wonderful and these goods as great. On its own, such bare recognition may not be enough to make these facts morally salient. However, what this recognition of the potential for wonderful future lives can still do is inspire us to want to bring about great goods in the future. And once we have been inspired to perform sacrifices that would contribute to bringing such futures into existence, then there are at least some beings who would be wronged, by rendering their sacrifices less worthwhile, if these futures are allowed to vanish along with our species.

References

Bader, Ralf (2022). The Asymmetry. In J. McMahan, T. Campbell, J. Goodrich and K. Ramakrishnan (eds.), *Ethics and Existence: The Legacy of Derek* Parfit. Oxford: Oxford University Press.

[10] But see (Bader 2022) for a rare exception.

[11] Such as Stanislav Petrov and Vasili Arkhipov, who disobeyed military orders to avert the firing of nuclear weapons during false alarms, or Bertrand Russell and other members of the Committee of 100, who sought to get arrested as a tactic to raise awareness of the risks form nuclear weapons.

[12] In response, one could say these things are absolutely good (or simply good). We don't find this move appealing, though others might. See especially Richard Kraut, who maintains that a thing can only be good for someone or good of a kind (Kraut 2011).

Beard, S.J. and Kaczmarek, Patrick (2019). On the Wrongness of Human Extinction, *Argumenta* 5(1): 85–97.

Chappell, Richard Yetter (2021). The Right Wrong-Makers. *Philosophy and Phenomenological Research* 103(2): 426–440.

Finneron-Burns, Elizabeth (2021). Human Extinction & Moral Worthwhileness: A Reply to Kaczmarek and Beard. *Utilitas* 34(1): 105–112.

Frick, Johann (2017). On the Survival of Humanity. *Canadian Journal of Philosophy* 47(2-3): 344–367.

Harman, Elizabeth (2009). Critical study: David Benatar. Better Never to Have Been: The Harm of Coming into Existence (Oxford: Oxford University Press, 2006). *Noûs* 43(4): 776–785.

Hurka, Thomas and Shubert, Esther (2012). Permission to Do Less Than the Best: A Moving Band. In M. Timmons (ed.), *Oxford Studies in Normative Ethics*, Volume 2 (pp. 1–27). Oxford: Oxford University Press.

Kaczmarek, Patrick and Beard, SJ (2020). Human Extinction and Our Obligations to the Past. *Utilitas* 32(2): 199–208.

Kamm, Frances (1985). Supererogation and Obligation. *Journal of Philosophy* 82(3): 118138.

Kraut, Richard (2011). *Against Absolute Goodness*. Oxford: Oxford University Press.

Lazar, Seth (2019). Accommodating Options. *Pacific Philosophical Quarterly* 100(1): 233–255.

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

Pritchard, H. A. (1965). Does Moral Philosophy Rest on a Mistake? *Mind* 21(81): 21–37.

Sider, Theodore (1993). Asymmetry and Self-Sacrifice. *Philosophical Studies* 70(2): 117–132.

Stocker, Michael (1976). Agent and Other: Against Ethical Universalism. *Australasian Journal of Philosophy* 54(3): 206–220.

Emil Andersson[1], Gustaf Arrhenius[2] & Tim Campbell[3]

Scanlonian Contractualism and Future Generations

In this paper we shall consider problems in population ethics for Scanlonian Contractualism. As we shall see, there are features of this view that make it difficult for it to satisfy rather obvious intuitive desiderata in population ethics. Rahul Kumar has suggested that his idea of "standpoints" offers help with some of these cases. We shall discuss different interpretations of this idea, and argue that it unfortunately fails. Scanlonian Contractualism cannot, it seems, avoid the aggregation problems that standard "impersonal" theories face without running into other problems that are at least as troublesome.

[1] Institute for Futures Studies, Department of Philosophy, McGill University, emil.andersson@iffs.se

[2] Institute for Futures Studies, Department of Philosophy, Stockholm University, Gustaf.Arrhenius@iffs.se, https://www.iffs.se/en/research/researchers/gustaf-arrhenius/

[3] Institute for Futures Studies, timothy.campbell@iffs.se, https://iffs.academia.edu/TimCampbell

1. A Challenge for Scanlonian Contractualism

Consider the following two populations:

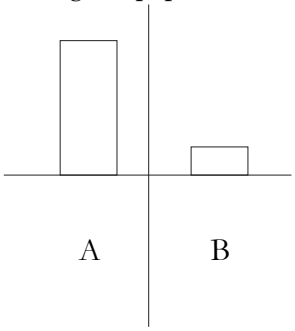


Diagram 1 Two Futures

In Diagram 1, the width of each block represents the number of people whereas the height represents their lifetime welfare. A population could consist of all the past, present and future lives in a possible world, or all the present and future lives, or all the lives during some shorter time span in the future such as the next generation, or all the lives that are causally affected by, or consequences of a certain action or series of actions, and so forth.[4]

All the lives in the diagram have positive welfare, or, as we also could put it, all the people have lives worth living. The A-people have very high positive welfare whereas the B-people have very low positive welfare.[5] The reason for this could be that in the B-lives there are, to paraphrase Derek Parfit, only enough ecstasies to just outweigh the agonies, or that the good things in those lives are of uniformly poor quality, e.g., eating potatoes and listening to Muzak.[6]

Depending on what we do, different people will exist in the future. We shall here assume, if not otherwise indicated, that different people exist in A and B. Which

[4] More exactly, a population is a finite set of lives in a possible world. A, B, C,... A_1, A_2,..., A_n, $A \cup B$, and so on, denote populations of finite size. We shall adopt the convention that populations represented by different letters, or the same letter but different indexes, are pairwise disjoint. For example, $A \cap B = A_1 \cap A_2 = \varnothing$.

[5] We shall say that a life has *neutral welfare* if and only if it is equally as good for the person living it as a neutral welfare component, and that a life has *positive* (*negative*) welfare if and only if it has higher (lower) welfare than a life with neutral welfare. A welfare component is neutral relative to a certain life *x* if and only if *x* with this component has the same welfare as *x* without this component. There are a number of alternative definitions of a neutral life in the literature, many of which would also work fine in the present context. For a discussion, see Arrhenius (2000), (forthcoming) ch. 2 and 9, Broome (1999), (2004), Bykvist (2007), p. 101, and Parfit (1984), pp. 357–358 and appendix G. Notice also that we actually don't need an analysis of a neutral welfare in the present context but rather just a criterion, and the criterion can vary with different theories of welfare.

[6] See Parfit (1984), p. 388 and Parfit (1986), p. 148.

future is better? Which is the one that we ought to aim for? It seems obvious that A is better; that is, the population with a very high quality of life and perfect equality is better than a same-sized population with a much lower quality of life. The claim that A is better than B also follows from what is probably the most uncontested adequacy condition in population ethics:

> *Egalitarian Dominance*: If 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.[7]

It also seems obvious that A is the population that we ought morally to bring about when faced with a choice between A and B. The claim that we ought morally to choose A when faced with a choice between A and B, and that it would be wrong to choose B, follows from a normative version of the Egalitarian Dominance Condition:

> *Normative Egalitarian Dominance*: 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, in any situation involving a choice between A and B, it is wrong to choose B, other things being equal.

This condition is, we think, as plausible as its axiological counterpart. The ceteris paribus condition involved here is a natural extension of the ceteris paribus condition used in the discussion of different axiologies:[8] There are neither any constraints (for example, promise-keeping) nor options (for example, great personal sacrifice for the agent which is beyond the call of duty), nor any non-welfarist values in the outcomes (for example, cultural diversity) that give us a reason to (not) choose one or the other of the involved actions. The only reasons for choosing one or the other of the involved actions arise from the welfare of the lives in the involved populations.

Consider a situation where you could, at no cost to yourself (you might even be among the beneficiaries), and without violating any other duties or compromising any other values, choose an outcome A in which everybody is equally well off, and better off as compared to another outcome B involving the same number of people. Surely it would be wrong to choose the latter outcome in this situation.

One might object here that whether it is morally wrong to choose B depends on

[7] The *ceteris paribus* clause in the formulation is meant to imply that the compared populations are roughly equal in all other putatively axiologically relevant aspect apart from individual welfare levels.

[8] See Arrhenius (2000), (forthcoming), see section 3.3.

the welfare level of its members.[9] If all the lives in this outcome have a very high welfare level, then perhaps choosing outcome A would be supererogatory. However, since these actions don't involve any kind of personal sacrifice for the agent, they don't fit the paradigm description of supererogatory actions, such as someone rushing into a burning house to save its residents at the risk of her own life, or someone donating a great sum of money to charity.[10] Moreover, it is hard to see any reason for why it should be optional to choose B in the cases that falls under the condition's domain. Hence, we think this objection misses its target.

Nevertheless, for our arguments below, and for the evaluation of the populations in Diagram 1, we could employ a weaker version of Normative Egalitarian Dominance in which the B-people have at most very low positive welfare. With this revision, it is even harder to find a reason for considering the choice of the outcome in which everyone is better off a supererogatory action. However, since we didn't find the objection persuasive in the first place, we shall keep the above formulation of the condition.

One might think that there is a way in which other things cannot be equal since at least one of the actions in a choice situation will be an omission and one might think that this is of relevance for an action's moral status. We are not convinced that the former claim is true since, arguably, an action is an "active" act or omission under a certain description and it is open to us to describe all the actions in the situations we are going to consider as acts rather than omissions. At any rate, even if it is true that at least one action in a situation is bound to be an omission, there are clear cases in which this property doesn't affect an action's deontic status such as when the consequence of the omission is much worse than that of the other alternatives. Consequently, we could restrict the conditions presented here to only apply to comparisons between "active" actions and not to cover omissions and then include in the situation under consideration a very bad "omission alternative" that is forbidden anyway.

We shall here consider the question of which future we should aim for from the point of view of Scanlonian Contractualism. According to this view, an act is morally right if and only if, and because, it is permitted by a principle that no one could reasonably reject. Otherwise, the act is morally wrong.[11] More specifically, Scanlonian Contractualism offers a criterion of moral rightness and wrongness narrowly construed as what people owe to each other. This excludes certain impersonal considerations that might matter within morality construed more broadly. Unsurprisingly, we find it utterly plausible that in a choice between populations A and B, not only is it true that we ought to bring about A, but to bring about A is a requirement of moral-

[9] Arrhenius is grateful to Michael Zimmerman for pressing this point.

[10] See Heyd (2008).

[11] Scanlon (1998), p. 4.

ity as narrowly construed by Scanlon. It is a requirement of what we owe to people. If not, we think that Scanlonian Contractualism has to be rejected, along with any morality that violates Normative Egalitarian Dominance.[12]

It has been surprisingly hard, however, for Scanlonian contractualists to reach this conclusion about choices between populations such as A and B. When he first put forward his theory, Scanlon claimed that his view was well equipped to deal with future people who are affected by our actions. However, he admitted that it was "less clear how it can deal with the problem presented by future people who would not have been born but for actions of ours which made the conditions in which they live worse".[13] Scanlon thus appeared to think that his theory was unable to deal with the normative version of the non-identity problem, the version that considers what we ought to do when our decision will determine who will, or will not, exist. In his later work he seemed to have changed his mind but was rather coy on the issue. Rather than expressing doubts over whether his theory could handle the non-identity problem he claimed that it "is a substantive question about when we have wronged someone, not a question about who can be wronged".[14]

Despite these later remarks by Scanlon, it is in fact unclear how his view can handle non-identity cases. To see why, consider the element of his contractualism that is usually referred to as the Individual Reasons Restriction: A person can reasonably reject a principle only on the basis of personal reasons—i.e. those that refer exclusively to the consequences for that person of others acting in ways permitted by the principle.[15] This requirement entails that certain considerations, such as the aggregate value of outcomes or the combined force of two or more individuals' personal reasons, do not provide grounds for reasonably rejecting a principle. This feature of his view Scanlon considers to be of central importance, since it is, according to him, "what enables it to provide a clear alternative to utilitarianism and other forms of consequentialism".[16]

This central element of Scanlon's theory creates a problem regarding how to handle the Two Futures case. If the theory is to imply that it would be wrong to create population B, it must be the case that there is someone who has a personal reason of sufficient weight to reasonably reject B. For this to be the case, there must be someone who is personally burdened by this act. But the people in B are created with lives worth living, and the only alternative for them is non-existence. This suggests that they are not burdened by the creation of population B -- they are not

[12] Those who disagree with this claim will face a different kind of problem, which we discuss in Section 3.
[13] Scanlon (1982), p. 115 n.10.
[14] Scanlon (1998), p. 186.
[15] Scanlon (1998), pp. 220, 229–230.
[16] Scanlon (1998), p. 229.

made worse off than they otherwise would have been since the only alternative is non-existence–and hence cannot reasonably reject such an act. Since Scanlonian Contractualism does not allow any other grounds for reasonable rejection, and since there seems to be no one else who is burdened by the creation of population B rather than population A, the theory appears unable to support the claim that it would be morally wrong to choose A over B. Thus, it appears unable to support Normative Egalitarian Dominance.

2. Can Scanlonian Contractualism Do Better?

In his most recent work Scanlon suggests, though very hesitantly, a possible solution to the non-identity problem. It may be possible for his contractualism to account for the wrong involved in bringing about B rather than A, he now says, by recognizing that "the objections that are relevant in the process of contractualist justification are not objections of particular individuals", but rather those objections that "any individual would have in virtue of being in a certain position".[17] Would, then, the people who find themselves in the position of those in B – who live under very bad conditions, but for whom the alternative would be to not exist at all – have an objection to a principle that allows the choice of B over A? About this Scanlon says that it is "not obvious to me that people in this position do not have such an objection, although I do not have a worked out view of the matter".[18]

As Scanlon admits that he has not worked out the details of this kind of solution, it is not surprising that he stops short of claiming that it will in fact be successful. But there are others who have already attempted to work out the details of the view that Scanlon appears to have in mind. In a series of influential papers, Rahul Kumar has suggested that Scanlonian Contractualism can be rendered immune to the normative Non-Identity Problem by conceiving of what is owed to particular individuals as being dependent on what is owed to them as "types", or as occupying certain "standpoints". A standpoint, as Kumar understands it, is not a determinate person. It is instead "a way of referring to a cluster of normatively significant characteristics (and related interests) that may aptly characterize certain actual particular individuals in actual situations in which they find themselves".[19] An example of what Kumar has in mind is provided by the case of a couple who are to conceive a child. What this couple owes to their future child does not depend on the particular psycho-physical identity of the particular person who will turn out to be their child.

[17] Scanlon (2021), p. 143.

[18] Scanlon (2021), p. 143.

[19] Kumar (2009), p. 261.

They owe a certain kind of consideration to whoever will instantiate the standpoint of their future child.[20]

We can now apply Kumar's suggestion to the case Two Futures. Even though there will be different people existing in A and in B, there need not be different standpoints. Since this is a same-number case there is a fixed number of possible standpoints, which will be occupied by the people in A or the people in B, or so we can assume (so for any standpoint in A, there is a corresponding standpoint in B – more on this below). Just as there is the standpoint of a couple's future child, there are the standpoints of future people; we need to consider what is reasonably rejectable from the point of view of the relevant standpoints, not the point of view of the particular persons who will turn out to occupy them.

More exactly, we can say:

> *Standpoint Contractualism 1 (SC1)*: If, for two populations A and B, each standpoint in A has a corresponding standpoint in B so in a bijection from A to B, every standpoint is better off in A than in B, and thus each standpoint can complain if B rather than A came about, then A is the right choice.

However, it might be unnecessary to compare the situation for each standpoint in A and B. Determining what is reasonable to reject involves a comparison of possible objections to proposed principles. But since there is no aggregation of personal objections, we only need to identify those standpoints that would be the most burdened by the relevant alternatives. A principle that no one can reasonably reject is thus, as Kumar puts it, "the principle whose implications are the most acceptable to the person to whom it is least acceptable".[21] An alternative to SC1 would then be:

> *Standpoint Contractualism 2 (SC2)*: If the standpoint(s) with the biggest potential welfare loss(es), and thus the biggest complaint(s), in the choice between two populations A and B, are better off in A than in B, then A is the right choice.

In Diagram 1, every standpoint has the same potential loss if B came about rather than A so SC2 will imply that A is the right choice.

Both SC1 and SC2 yield the correct verdict in the case depicted in Diagram 1. But while this case is a different-people case, it is greatly simplified by being a same-number case. In different-number cases, the focus on standpoints runs into trouble.

[20] Kumar (2003), pp. 113–114.

[21] Kumar (2018), p. 248.

Consider now:

> *The Repugnant Conclusion*: For any population consisting of people with very high positive welfare, there is a better population in which everyone has a very low positive welfare, other things being equal.[22]

Many find this conclusion implausible. Moreover, there is a normative version of it, what we might call the Normative Repugnant Conclusion, according to which for any population consisting of people with very high positive welfare, there is a population in which everyone has very low positive welfare, such that in a choice between the former and the latter, one ought morally to choose the latter, other things being equal. The two populations described in the above statement of the Repugnant Conclusion are labelled A and Z in Diagram 2:

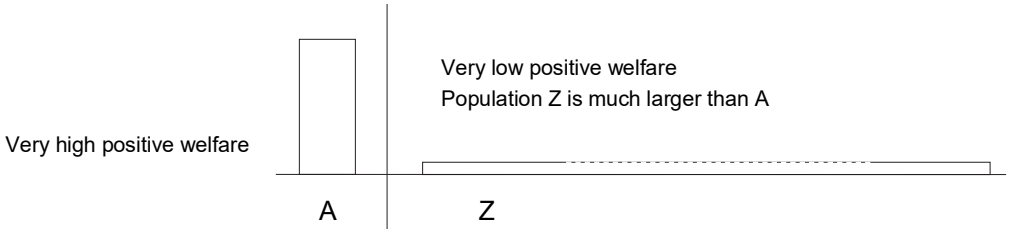


Diagram 2 The Repugnant Conclusion

SC1 is not applicable to the comparison of A and Z, since there is no bijection from A to Z (we will get back to SC2's implication in this case). We have a lot more people in Z than in A. Thus, we have a lot of new standpoints. Ideally, we want a version of SC that can not only accommodate Normative Egalitarian Dominance, but also avoid the Normative Repugnant Conclusion.

One suggestion is:

> *SC3*: As SC1 but we only look at the standpoints that exist in both of the compared populations and ignore all the other standpoints (and hence reduce the comparison to a same-number case).

[22] Here's how Parfit (1984), p. 388 formulates the conclusion: "For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better, even though its members have lives that are barely worth living." Hence, our formulation is more general than his.

SC3 will give the right result and yield that A is the right choice in Two Futures (Diagram 1). It would also deliver the result that we ought morally to choose A over Z in Diagram 2, provided that we assume that the standpoints in A also exist in Z.

So far, so good. However, consider:

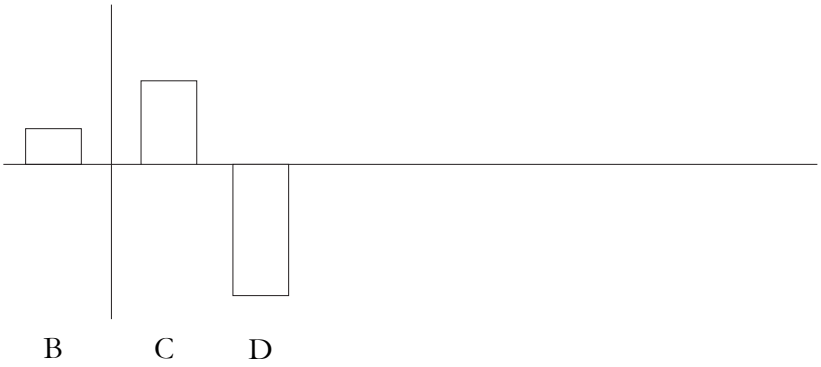


Diagram 3

In Two Futures, we saw that Kumar's view is compatible with the claim that the people in A and B occupy the same standpoints. But in cases such as the one depicted in Diagram 3 it is not clear which standpoints overlap. Does B overlap with C, with D, or with some portions of both? The answer to this question is crucial, as it determines the implications of SC3. If we assume that at least some of the standpoints in B overlap with some of the standpoints in D, SC3 will yield that B is the right choice. But if we instead assume that the standpoints in B correspond to the standpoints in C, SC3 will counterintuitively yield that C∪D is the right choice.

To avoid this embarrassing result, Kumar has to find a way of individuating standpoints that rules out that the standpoints in B correspond to those in C on some description of these populations. We see no non-arbitrary way to do this.

However, a revision of SC3 might handle this problem:

> *SC4*: We make a bijection between the worst off standpoints and ignore the rest of the standpoints that are left over in the bigger population.

SC4 would imply that B is the right choice since the bijection of B onto D shows a big complaint from the D-standpoints if C∪D were picked.

Let us get back to SC2's implications. Since it focuses on the biggest complaints,

it also yields that B is the right choice above. Moreover, it also yields the right answer to the Repugnant Conclusion in Diagram 2.

However, a problem with individuation of standpoints arises here. As we noted above, Kumar hasn't provided any sharp criterion for individuation of standpoints. But he does claim that psycho-physical identity is "of no moral relevance". This suggests that the fact that the same person exists in either of two compared outcomes is neither necessary nor sufficient for correspondence of standpoints.[23] Given this, we simply cannot say whether the B- and C-people in Diagram 3 occupy the same standpoints. Hence, SC2 and SC4 do not get any support from the idea of standpoints and cannot be taken to be a revision of Scanlonian Contractualism in terms of standpoints.

There are further problems for SC2 and SC4:

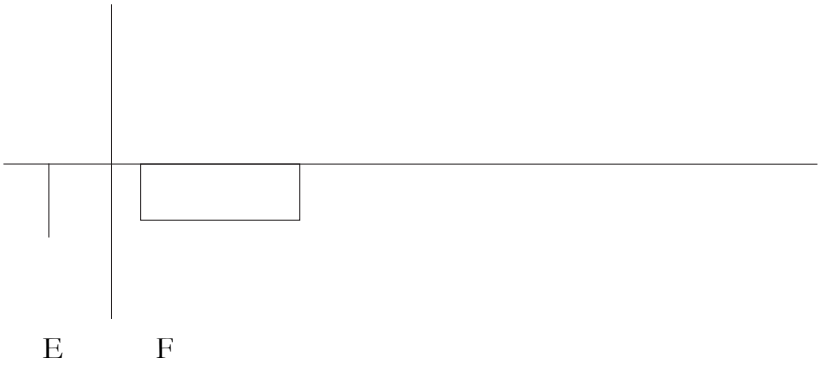


Diagram 4

In Diagram 4, all the involved people have negative welfare and suffer, and just slightly less so in F as compared to E. SC2 and SC4 imply that F is the right choice, irrespective of how many suffering people there are in F and how much they suffer. Given either SC2 or SC4, there is a description of E and F according to which there are corresponding standpoints in E and F. SC2 and SC4 would imply that F is the right choice even if only one person in F was better off than the E-person (and assuming a correspondence of the standpoints occupied by the one F-person and the E-person). Moreover, if all in F were as badly off as the E-person, SC2 and SC4 would imply that it is right to choose either E or F.

One solution here would be to count the welfare losses of standpoints in both of

[23] For discussion of a related point, see Valeska Martin (2022).

the compared populations, but then also take into account the welfare of uniquely realisable standpoints, that is, standpoints that only exist in one of the two compared populations. Basically, we compare the welfare of each uniquely realisable standpoint in one population to its non-existence in the other; if the standpoint has negative welfare then its existence counts as a relevant welfare loss, and if it has positive welfare then its non-existence counts as a relevant welfare loss. But then we are in the aggregation game and will get all the aggregation problems that standard "impersonal" theories get. For example, if we assume that it is better for a person to exist than not to exist, then we would be led to the Normative Repugnant Conclusion.

Another solution would be to abandon the assumption that the number and correspondence between standpoints must be determined strictly by the number of people in different populations. Perhaps when we are creating different numbers of future people, there is but one standpoint that we can denote with the term 'future person'. On this proposal, there is, for example, only one relevant standpoint in E and only one relevant standpoint in F, and there is a correspondence between them. Standpoints are "elastic" in the sense that any positive number of people can occupy a single standpoint.

However, we will then need to say how to aggregate welfare within a standpoint. One possibility is that the welfare within a standpoint is the total welfare of the individuals that occupy it. But that takes us back to the Normative Repugnant Conclusion. Another possibility is that the welfare within a standpoint is the average welfare of the individuals that occupy it. But that leads to the crazy conclusion that it is right to choose F in a choice between E and F. Basically, on this "elastic standpoint" view, we end up with the same aggregation problems that standard "impersonal" theories get, accept that these problems will be exemplified within, rather than across, standpoints. We don't think that this difference makes these aggregation problems any more palatable.[24]

3. Pluralism

We have argued that Scanlonian Contractualism either cannot explain the wrongness of an act or policy that causes people to exist who are substantially worse off than others who would have existed under some alternative act or policy, or that it leads to the same kinds of aggregation problems the avoidance of which motivates one of its central features—the Individual Reasons Restriction. All the while, we

[24] We thank Per Algander for pointing out that Kumar's view can be interpreted in terms of "elastic" standpoints.

have assumed that contractualism ought to accommodate Normative Egalitarian Dominance, and that otherwise, it should be rejected.

However, it is worth considering what happens if the contractualist denies this assumption. In a recent paper addressing the non-identity problem, Scanlon considers the possibility of a pluralistic morality that views the non-identity problem as being outside the purview of what we owe to each other, and hence, outside the scope of his contractualism.[25] According to Scanlon, this pluralist response to the non-identity problem concedes "that contractualism cannot explain the wrongfulness of a policy that makes people who live at some future time much worse off than the people who would have lived at that time if some alternative policy had been followed" but explains the wrongness of the policy "simply by the fact that it brings about a situation that is much worse than the situation produced by such an alternative" (2021, 142).

The question is how this moral pluralism should work. One possibility is that consequentialist considerations are relevant only in the non-identity cases, or perhaps in a wider range of cases in which our decision will not make anyone worse off than they would otherwise be.

But the problem is that there is no reason to believe this. If the fact that one outcome is much worse than another is morally relevant in the non-identity cases, then it is also morally relevant in same-person cases. This means that consequentialist reasons will be relevant even in cases that Scanlon would want to say are given an attractive treatment by his contractualism. But it is hard to see how the consequentialist and contractualist reasons can be combined in a way that avoids the problems that each of these theories face when considered separately. Suppose that the consequentialist and contractualist reasons can be compared and weighed against each other. Then it seems likely that the consequentialist reasons will often spoil Scanlon's preferred contractualist treatment of certain cases, such as cases in which we can either give a huge benefit to one person or a tiny benefit to many others. Since interpersonally aggregated goodness is relevant even in these cases, there will be some number of tiny benefits that outweighs the huge benefit, even when both consequentialist and contractualist reasons are taken into account.

4. Summary

We have argued that Scanlonian Contractualism cannot accommodate Normative Egalitarian Dominance, and hence, cannot offer a plausible treatment of the part of morality that concerns future generations. If the non-identity problem and other

[25] Scanlon (2021).

problems concerning future generations are deemed outside the scope of Scanlonian Contractualism, and consequentialist considerations are brought in to deal with these problems, then these consequentialist considerations will be relevant even in the same-person cases where Scanlonian Contractualism is relevant. A pluralism that incorporates elements of the two different moral theories is therefore problematic, as there is no apparent way of combining the consequentialist and Contractualist elements in a way that Scanlonian Contractualists would find acceptable. Moreover, pluralism runs the risk of inheriting what is deeply problematic in both Scanlonian Contractualism and consequentialism, rather than solving the problems that each theory faces.[26]

References

Arrhenius, G. (2000*). Future Generations: A Challenge for Moral Theory*. Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2:170236

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

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

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

Bykvist, K. (2007). The Good, The Bad, and the Ethically Neutral. *Economics and Philosophy*, 23(01), 97–105. https://doi.org/10.1017/S0266267107001253

Heyd, D. (2008). Supererogation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (pp. 1–23). Retrieved from http://plato.stanford.edu.ezp.sub.su.se/archives/win2012/entries/supererogation

Kumar, R. (2003). Who Can Be Wronged? *Philosophy and Public Affairs*, 31(2), 99–118.

Kumar, R. (2009). Wronging Future People: A Contractualist Proposal. In A. Gosseries & L. H. Meyer (Eds.), *Intergenerational Justice*. Oxford: Oxford University Press.

[26] Financial support from Riksbankens Jubileumsfond (grant M17-0372:1) is gratefully acknowledged. We would like to thank Per Algander, Krister Bykvist, Valeska Martin, Daniel Ramöller, and participants at the 2020 workshop Contractualism and Future Generations at the Institute for Futures Studies, Stockholm.

Kumar, R. (2018). Risking Future Generations. *Ethical Theory and Moral Practice*, 21(2), 245–257.

Valeska Martin, D. (2022). Moral Contractualism, Non-Identity, and Types of Persons. Unpublished Manuscript.

Parfit, D. (1984). *Reasons and Persons* (1991st ed.). Oxford: Clarendon.

Parfit, D. (1986). Overpopulation and the Quality of Life. In P. Singer (Ed.), *Applied Ethics* (1 edition, pp. 145–164). New York: Oxford University Press.

Scanlon, T.M. (1982). Contractualism and Utilitarianism. In B. A. O. Williams & A. Sen (Eds.), *Utilitarianism and Beyond* (pp. 103–128). Cambridge: Cambridge University Press.

Scanlon, T.M. (1998). *What We Owe to Each Other*. Cambridge, Mass.: Belknap.

Scanlon, T.M. (2021). Responses to Forst, Mantel, Nagel, Olsaretti, Parfit, and Stemplowska. In M. Stepanians & M. Frauchiger (Eds.), *Reason, Justification, and Contractualism*: Themes from Scanlon (pp. 131-153). Boston: Walter de Gruyter GmbH.

Studies on climate ethics and future generations, vol. 1 Working paper series 2019:1–11. Eds. Paul Bowman & Katharina Berndt Rasmussen

Tim Campbell: *The Bullet-Biting Response to the Non-Identity Problem*

Melinda A. Roberts: *Does the Additional Worth-Having Existence Make Things Better?*

Anders Herlitz: *Nondeterminacy and Population Ethics*

Wlodek Rabinowicz: *Can Parfit's Appeal to Incommensurabilities Block the Continuum Argument for the Repugnant Conclusion?*

Gustaf Arrhenius & Julia Mosquera: *Positive Egalitarianism*

Marc Fleurbaey & Stéphane Zuber*: Discounting and Intergenerational Ethics*

Stéphane Zuber: *Population-Adjusted Egalitarianism*

Katie Steele: *'International Paretianism' and the Question of 'Feasible' Climate Solutions*

Göran Duus-Otterström: *Sovereign States in the Greenhouse: Does Jurisdiction Speak against Consumption-Based Emissions Accounting?*

Paul Bowman: *On the Alleged Insufficiency of the Polluter Pays Principle*

Martin Kolk: *Demographic Theory and Population Ethics – Relationships between Population Size and Population Growth*

Studies on climate ethics and future generations, vol. 2 Working paper series 2020:1–11. Eds. Paul Bowman & Katharina Berndt Rasmussen

Krister Bykvist: *Person-affecting and non-identity*

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

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

Krister Bykvist & Tim Campbell: *Persson's Merely Possible Persons*

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

Paul Bowman: *Duties of Corrective Justice and Historical Emissions*

Katie Steele: *The distinct moral importance of acting together*

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

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

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

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

Studies on climate ethics and future generations, vol. 3 Working paper series 2021:1–10. Eds. Joe Roussos & Paul Bowman

Dean Spears & H. Orri Stefánsson: *Calibrating Variable-Value Population Ethics*

Joe Roussos: *Awareness Growth and Belief Revision*

Katie Steele: *Why Time Discounting Should Be Exponential: A Reply to Callender*

Nicholas Lawson & Dean Spears: *Population Externalities and Optimal Social Policy*

John Broome: *How Much Harm Does Each of Us Do?*

Tim Campbell: *Offsetting, Denialism, and Risk*

Paul Bowman: *The Relevance of Motivations to Wrongdoing for Contributing to Climate Change*

Hilary Greaves & John Cusbert: *Comparing Existence and Non-Existence*

M.A. Roberts: *Does Climate Change Put Ethics on a Collision Course with Itself?*

Anders Herlitz: *Fixing Person-Based Stakes in Distributive Theory*

Studies on climate ethics and future generations, vol. 4
Working paper series 2021:11–23. Eds. Joe Roussos & Paul Bowman

Gustaf Arrhenius: *Democratic Representation of Future Generations and Population Ethics*

Elizabeth Finneron-Burns: *Global Justice and Future Generations: The Case of Sovereign Wealth Funds*

Paul Bou-Habib & Serena Olsaretti: *Children or Migrants as Public Goods?*

Tim Campbell, Martin Kolk & Julia Mosquera: *Universal Procreation Rights and Future Generations*

H. Orri Stefánsson: *What Is the Point of Offsetting?*

Göran Duus-Otterström: *The Role of Subsistence Emissions in Climate Justice*

Julia Mosquera: *Climate Change, Corrective Justice, and Non-Human Animals*

Göran Duus-Otterström & Edward A. Page: *Defeating Wrongdoing: Why Victims of Unjust Harm Should Take Priority over Victims of Bad Luck*

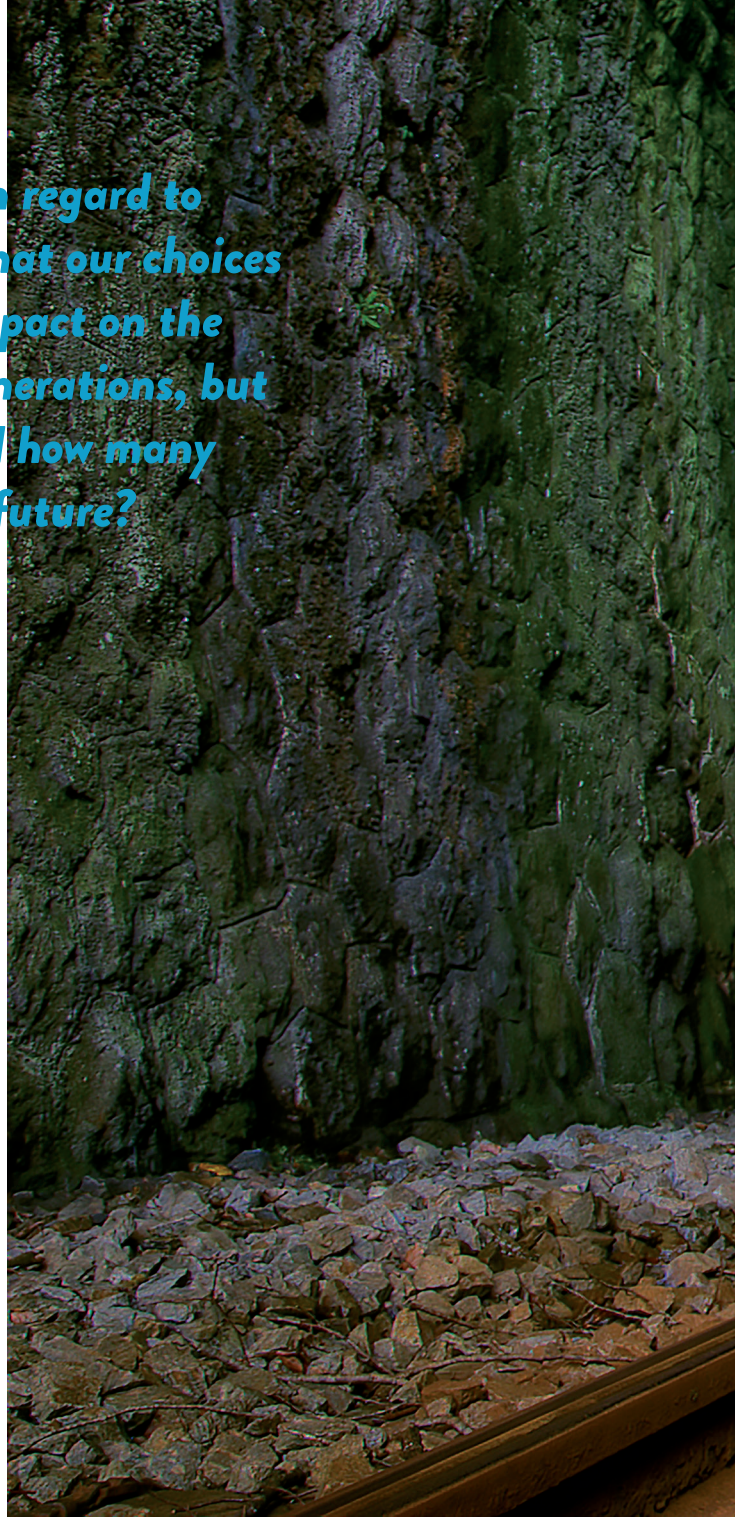
Henrik Andersson, Eric Brandstedt & Olle Torpman: *Review Article: The Ethics of Population Policies*

Partha Dasgupta & S.J. Beard: *Optimum Population and Environmental Constraints – A Utilitarian Perspective*

M.A. Roberts: *Anonymity and Indefinitely Iterated Addition and Reversal*

Henrik Andersson & Anders Herlitz: *Classifying Comparability Problems in a Way that Matters*

Krister Bykvist & Tim Campbell: *Frick's Defence of the Procreation Asymmetry*



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 is the fifth volume in our series. It collects eleven contributions from researchers in philosophy, social psychology and demography.

Find more information at climateethics.se.

INSTITUTE FOR FUTURES STUDIES

Box 591, SE-101 31
Stockholm, Sweden

Phone:
+46 8 402 12 00

E-mail:
info@iffs.se

Institute for Futures Studies