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

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

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

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

Find more information at climateethics.se.

INSTITUTE FOR FUTURES STUDIES

Box 591, SE-101 31
Stockholm, Sweden

Phone:
+46 8 402 12 00

E-mail:
info @iffs.se

Institute for Futures Studies

Vol. 3

STUDIES ON CLIMATE ETHICS AND FUTURE GENERATIONS

Editors: Joe Roussos & Paul Bowman

WORKING PAPER SERIES
Vol. 3
2021:1-10

STUDIES ON CLIMATE ETHICS AND FUTURE GENERATIONS

Working paper series 2021:1–10

Editors: Joe Roussos & Paul Bowman

Studies on Climate Ethics
and Future Generations
Vol. 3

Studies on Climate Ethics and Future Generations Vol. 3

Editors: Joe Roussos
& Paul Bowman

Institute for Futures Studies
Working Papers 2021:1-10
Stockholm 2021

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

Cover: Matilda Svensson
Cover image: Getty images
Print: Scandbook, Falun, Sweden, 2021
Distribution: The Institute for Futures Studies, 2021

FSC
www.fsc.org
MIX
Paper from
responsible sources
FSC® C021394

Contents

Preface 7

Calibrating Variable-Value Population Ethics
Dean Spears & H. Orri Stefánsson 9

Awareness Growth and Belief Revision
Joe Roussos 29

Why Time Discounting Should Be Exponential: A Reply to Callender
Katie Steele 65

Population Externalities and Optimal Social Policy
Nicholas Lawson & Dean Spears 81

How Much Harm Does Each of Us Do?
John Broome 111

Offsetting, Denialism, and Risk
Tim Campbell 125

The Relevance of Motivations to Wrongdoing for Contributing to Climate Change
Paul Bowman 137

Comparing Existence and Non-Existence
Hilary Greaves & John Cusbert 163

Does Climate Change Put Ethics on a Collision Course with Itself?
M.A. Roberts 197

Fixing Person-Based Stakes in Distributive Theory
Anders Herlitz 221

Preface

This is the third volume of preprint papers produced by the Climate Ethics and Future Generations project. The project is led by PI Gustaf Arrhenius and co-PIs Krister Bykvist and Göran Duus-Otterström and hosted by the Institute for Futures Studies in Stockholm. It is interdisciplinary, including work in philosophy, political science, psychology, sociology, and economics. The project, which runs from 2018–2023, aims to provide comprehensive and cutting-edge research on ethical questions concerning future generations in the context of climate change policy. Climate Ethics and Future Generations is generously financed by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences). For more information, visit climateethics.se.

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

This volume demonstrates particularly well our interdisciplinary work combining economics and philosophy. First is a joint paper by Dean Spears (an economist) and H. Orri Stefánsson (a philosopher), in which they study a family of "variable-value" population axiologies largely developed by economists. Spears and Stefánsson argue that these views imply conclusions which ought to be as repugnant as the so-called Repugnant Conclusion

Joe Roussos and Katie Steele investigate questions of rationality that lie on the boundary of philosophy and economics. Roussos looks at how rational agents should change their beliefs when they encounter unforeseen possibilities. He defends a "Reverse Bayesian" theory of awareness growth from two recent objections. Steele looks at the question of rational time-discounting. She argues that rational agents ought to have exponential time-discounting functions.

Next, Nicholas Lawson and Dean Spears study an economic question in optimal social policy which is motivated by the philosophical issue of accounting for future generations. They examine the interaction between the education of (potential) parents and fertility, in terms of their population-level effects. John Broome, himself an economist and philosopher, is the author of our fifth paper. Broome offers a largely empirical effort to account for how much harm individuals do via their emissions, by examining in detail the results of a report titled "Valuing the Global Mortality Consequences of Climate Change." In his contribution, Tim Campbell identifies a tension between Broome's view that an individual causes serious harm with their emissions and his view (from an earlier work) that individuals can satisfy their duty not to cause harm by offsetting their emissions. Campbell seeks to resolve this tension by arguing that the relevant duty to reduce one's carbon footprint is the duty to limit risk, rather than the duty to avoid causing harm. For his part, Paul Bowman addresses related questions of individual wrongdoing for collectively-caused harms, arguing that an individual's motivations when contributing to such harms can alter the nature and extent of their wrongdoing.

The next two papers in the volume discuss questions in the foundations of population ethics, relating to the value of coming into existence. Hilary Greaves and John Cusbert defend a controversial thesis about the comparative value of existence. They claim that it is not incoherent to hold that it can sometimes be better or worse, for a given person, that that person exist rather than not. Melinda Roberts is similarly concerned with the value of existence. She defends her formulation of the person-affecting view from two objections, each of which turns on the claim that probabilities are, in at least some cases, critical to moral evaluation.

Our final paper, by Anders Herlitz, addresses views of distributive justice which hold that distribution should be in accordance with what is at stake for the individuals involved. Herlitz argues that stakes-based proposals either fail to meet requirements of rationality or are incompatible with the theories of justice they are intended to serve.

We are pleased to be able to share this selection of work from the Climate Ethics and Future Generations project. We encourage readers to contact the authors with comments, questions, and objections—making this volume a part of the work of our project.

Joe Roussos & Paul Bowman
Editors

Dean Spears[1] & H. Orri Stefánsson[2]

Calibrating Variable-Value Population Ethics[3]

Variable-Value axiologies propose solutions to the challenges of population ethics. These views avoid Parfit's *Repugnant Conclusion*, while satisfying some weak instances of the *Mere Addition* principle (for example, at small population sizes). We apply calibration methods to Variable-Value views while assuming: first, some very weak instances of Mere Addition, and, second, some plausible empirical assumptions about the size and welfare of the intertemporal world population. We find that Variable-Value views imply conclusions that should seem repugnant to anyone who opposes Total Utilitarianism due to the Repugnant Conclusion. So, any wish to avoid repugnant conclusions is not a good reason to choose a Variable-Value view. More broadly, these calibrations teach us something about the effort to avoid the Repugnant Conclusion. Our results join a recent literature arguing that prior efforts to avoid the Repugnant Conclusion hinge on inessential features of the formalization of repugnance. Some of this effort may therefore be misplaced.

[1] Economics Department and Population Research Center, University of Texas at Austin; Economics and Planning Unit, Indian Statistical Institute - Delhi Centre; IZA; Institute for Future Studies, Stockholm; r.i.c.e. (www.riceinstitute.org). dean@riceinstitute.org

[2] Department of Philosophy, Stockholm University; Swedish Collegium for Advanced Study, Uppsala; Institute for Future Studies, Stockholm. orri.stefansson@philosophy.su.se

[3] We are grateful for very useful written comments from Jake Nebel and Christian Tarsney, and for helpful suggestions and questions from the audience at Climate Ethics Workshop, Institute for Futures Studies in Stockholm, April 2020. Although this paper received no specific funding, Spears' research is supported by NICHD grants K01HD098313 and P2CHD042849. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Stefánsson's research is supported by Riksbankens Jubileumsfond through a *Pro Futura Scientia* fellowship. Funding from Riksbankens Jubileumsfond to the Climate Ethics program is also gratefully acknowledged.

Introduction

Much research in population ethics — as studied by both philosophers and economists — is motivated by the quest to avoid what (Parfit 1984) called the *Repugnant Conclusion*, one version of which states that:[4]

> **The Repugnant Conclusion** (Informal version). For any perfectly equal population of very well-off people, there is a better population consisting entirely of lives that are barely worth living.

Total Utilitarianism, according to which a population is better the greater sum of welfare it contains, is widely recognized to entail the Repugnant Conclusion. No matter how well-off people are in some population A, and independently of A's size, there is some (potentially much bigger) imaginable population Z that contains a greater sum of welfare than A does — even though people in Z have lives that are each barely worth living (understood as having barely positive welfare).

Most paths to avoiding the Repugnant Conclusion begin by abandoning what Parfit called the *Mere Addition principle,* which can be stated thus:

> **Mere Addition** (Informal version). By adding any life worth living to any population, without making anyone else worse off, we do not make the population worse.

Total Utilitarianism implies the Mere Addition principle. But this principle is violated by *Average* Utilitarianism, according to which a population is better the greater average welfare it contains. And Average Utilitarianism avoids the Repugnant Conclusion: Population Z, whose members all have lives that are barely worth living, contains lower average welfare than A. So, A is better than Z, according to Average Utilitarianism.

Somebody who abandons Mere Addition argues that merely adding a life worth living, without making anyone worse off, can make a population worse. But what about adding a life *well* worth living? Consider merely adding a person who lives a very good life by modern standards: say, a professor living in a developed country in

[4] Parfit's own formulation of the Repugnant Conclusion states that: "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." (Parfit 1984, 388) Our formulation is closer to Arrhenius's (forthcoming). (Spears and Budolfson 2021) have argued that formalizations of the Repugnant Conclusion should be broader — including, for example, additions to unaffected, intersecting populations — but for this paper we ignore that proposal and focus on what they call a "restricted" formalization.

2020. Surely by adding a person like that to any population, without thereby making anyone else worse off, we have made the population better? Not according Average Utilitarianism. To see this in an absurd example: adding our professor to a single-person "population" whose member is only a tiny bit better-off than the professor makes the resulting population *worse*, according to Average Utilitarianism. In fact, if the future of humanity is as long and wonderful as some hope (Ord 2020), then adding a person likes this to the *actual* intertemporal world population makes the resulting population worse, according to Average Utilitarianism. This violates what we shall call *Weak Mere Addition* (which we state formally in section 2).

In light of the above counterintuitive implications of on the one hand Total Utilitarianism (Repugnant Conclusion) and on the other hand of Average Utilitarianism (violating Weak Mere Addition), some theorists have been attracted to a family of views that are often called *Variable-Value views*.[5] These views are intended to avoid the Repugnant Conclusion while capturing the intuition that adding a well-off person to a small population makes the resulting population better. More specifically, these views hold that the quantity that added persons (with a fixed level of welfare) contribute towards the overall value of a population decreases as the size of the population increases, *cumulatively* contributing only a bounded amount, which is how such views escape the Repugnant Conclusion.

Various versions of Variable-Value views have been rigorously formalized.[6] These formalizations and the ensuing analysis has focused on *qualitative* properties of Variable-Value population ethics: with which *axioms* do these proposals comply? However, there has not been a similar focus on the *quantitative* implications of these Variable-Value views. In particular, one might wonder *how fast* the quantity that an added person contributes towards the overall value of a population diminishes as e.g. the size and average welfare of the population increases, and what implications that will have for various trade-offs between increasing the size and the average welfare of a population. Similarly, one might wonder precisely which weakenings of the Mere Addition principle these views can accommodate without implying instances of the Repugnant Conclusion.[7]

[5] (Hurka 1983) coined the term, and was probably the first to suggest such a view in response to Parfit's Repugnant Conclusion, but views in this family have since been proposed or investigated by (Ng 1989), (Sider 1991), (Asheim and Zuber 2014), and (Pivato 2020), although not all of these authors endorsed the Variable-Value axiology that they identified or explored.

[6] Examples include (Ng 1989), (Asheim and Zuber 2014), and (Pivato 2020)

[7] Our aim is not to examine *all* Variable-Value views. In particular, we shall not be concerned with those variable-value views that satisfy the strong version of Mere Addition (i.e., the version entailed by Total Utilitarianism), such as the theory examined in Sider's (1991). Instead, the aim is to examine those views that (unlike Average Utilitarianism) imply some weak instance of Mere Addition, without implying the strong version of Mere Addition.

We note also that a normative reason for excluding from our examination the view in (Sider 1991) is that

Our aim in this paper is to fill the above gap in the population ethics literature. In particular, we shall assume some very weak instances of Mere Addition and calibrate what Variable-Value axiologies, that satisfy such weak instances of Mere Addition (but violate the stronger Mere Addition principle that Total Utilitarianism entails), imply under what we take to be plausible empirical assumptions about the future. Informally, the weak Mere Addition that we assume ensures that merely adding people who are very well-off by modern standards, such as professors in the developed world, does not make the population worse. The empirical assumption we make is that the future of humanity is long and prosperous, such that, in particular, the average welfare of the total intertemporal world population is higher than the average welfare of the world population up to 2020.[8]

Our main observation is that, when combined with the above two assumptions, Variable-Value axiologies imply countless instances of the Repugnant Conclusion. (By an "instance" of the Repugnant Conclusion, we mean the judgement that some particular population consisting only of lives that are barely worth living is better than some particular perfectly equal population of very well-off people.) Of course, they do not imply the *qualitative* Repugnant Conclusion stated above — which holds for *all* populations of very well-off people. But these implications, we argue, should nevertheless seem every bit as repugnant to those who oppose the Repugnant Conclusion.[9]

It might be worth providing some additional remarks to motivate our methodology.[10] First, we assume that even those who are happy with giving up the traditional Mere Addition principle will nevertheless find it hard to reject some very weak instances of the principle. After all, we seem to have stronger reasons to think that a mere addition of a very well off person does not make the world worse than we have to think that a mere addition of a life barely worth living does not make the world worse. Therefore, there is, we think, something to be gained from exploring what happens when we replace Mere Addition with a weaker principle that only applies to people who are very well-off.

Second, we think that valuable lessons can be learnt from exploring what population axiologies imply given reasonable empirical assumptions, as opposed to merely exploring what these axiologies imply in theory. In particular, our finding

it implies what Arrhenius's (forthcoming) calls "The Very Anti Egalitarian Conclusion: For any perfectly equal population of at least two persons with positive welfare, there is a population which has the same number of people, lower average (and thus lower total) welfare and inequality, which is better." In fact, Sider himself rejects the view due to implications like this (Sider 1991, 270).

[8] If the reader finds this empirical assumptions implausible, then she can of course read our argument and conclusion as being *merely conditional* on these assumptions.

[9] In fact, according to the principle of "unrestricted instantiation" (Tännsjö 2020), these implications *must* be seen as repugnant if the Repugnant Conclusion is to be an argument against Total Utilitarianism.

[10] Many thanks to Christian Tarsney for making us see the need to address the motivation.

that Variable-Value views have counterintuitive implications, given empirical assumptions that we accept for our actual world population, provides a valuable lesson that is not learnt from simply learning that these axiologies have counterintuitive implications given assumptions about the world population that we take to be false. For that shows that Variable-Value views do not only have counterintuitive implycations in hypothetical scenarios; they also haver counterintuitive implications in empirically plausible scenarios.

We proceed as follows. In Section 2 we lay out the formal framework of the paper, which allows us to state more formally the views and conditions we informally describe above, and introduce the reader to calibration methods in decision theory. Then, in Section 3, we use such methods to examine what two prominent Variable-Value views imply when they have been calibrated to the current world population and what we take to be reasonable assumptions about the future population. In Section 4 we use the same methods to present a more general result, that is, a result for all Variable-Value views that do not satisfy the strong version of Mere Addition. Finally, in Section 5 we ask what the upshot of our arguments is for population ethics and in particular for the focus in the population ethics literature on avoiding the Repugnant Conclusion.

Formal framework for population ethics

Our framework, terminology, and notation follow closely that of (Asheim and Zuber 2014). Let $\mathbb{N}$ denote the set of natural numbers and $\mathbb{R}$ the set of real numbers. Let $\mathbf{X} = \bigcup_{n\in\mathbb{N}} \mathbb{R}^n$ denote the set of possible finite distributions of lifetime well-being. More formally, $\mathbf{X} = \bigcup_{n\in\mathbb{N}} \mathbb{R}^n$ is a set of of vectors of real numbers, where each number represents the lifetime well-being of some person. A generic such vector for a population of m people is denoted $\mathbf{x} = (x_1, \dots, x_m)$, where x_i denotes the lifetime well-being of individual i. The size of the population given by $\mathbf{x}$ is denoted by $\mathcal{N}(\mathbf{x})$ (and will, as mentioned, always be finite). For any vector $\mathbf{x}$, we write the average lifetime well-being of its members as $\bar{x}$. So, $\bar{x}$ should be interpreted as the average lifetime welfare of people given by $\mathbf{x}$.

Built into our framework is an *anonymity* axiom, which holds that the "better-than relation" we study is invariant under permutations of the vectors in $\mathbf{X}$. So, for instance, let $\mathbf{x}'$ be the vector that results when the lifetime well-being of i and j in $\mathbf{x}$ are switched. Then the better-than relations that we shall consider are all indifferent between $\mathbf{x}$ and $\mathbf{x}'$, that is, they deem these two distributions to be equally good. Intuitively, this means that it does not matter *who* receives what welfare; all that matters is how many people have each welfare level. This assumption rules out some person-affecting views.

For any $\mathbf{x} \in \mathrm{X}$ with m members, let $\mathbf{x}_{[]} = (x_{[1]}, \ldots, x_{[r]}, \ldots, x_{[m]})$ be the nondecreasing reordering of $\mathbf{x}$. In other words, in $\mathbf{x}_{[]}$ the elements of $\mathbf{x}$ have been put in a nondecreasing order, such that for each rank $r \in \{1, \ldots, m\}$, $x_{[r]} \leq x_{[r+1]}$, meaning that individual with rank $r + 1$ is at least as well off as individual with rank r. The anonymity assumption ensures that when two or more individuals are equally well-off, how they are ranked relative to each other does not affect the ranking of populations.

Let $(z)_n \in \mathbb{R}^n$ denote the perfectly-equal distribution where all n individuals have lifetime well-being z. And let $(\mathbf{x}, (z)_n)$ denote distribution $\mathbf{x} \in \mathrm{X}$ with n added individuals that all have lifetime well-being z. When only one individual with well-being level y is added to $\mathbf{x}$, we denote this by $(\mathbf{x}, y)$.

Finally, $\precsim$ on X denotes a (weak) better-than relation on X, such that for any $\mathbf{x},\mathbf{y} \in \mathrm{X}$, $\mathbf{x} \precsim \mathbf{y}$ means that $\mathbf{y}$ is at least as good as $\mathbf{x}$. Throughout the discussion we shall assume that the better-than relation is transitive, reflexive, and complete,[11] which means that the relation generates a better-than *order*. The strict relation, $\prec$, and indifference, $\sim$, are respectively the asymmetric and symmetric counterparts of $\precsim$.

With this formalization, different axiological views, such as those discussed above, can be seen as different views about the structure of $\precsim$. This allows for convenient formal statements of the views and conditions we informally discussed in the last section. For instance, Total Utilitarianism can be formulated thus:

Total Utilitarianism (TU). *For any* $\mathbf{x},\mathbf{y} \in \mathrm{X}$*:*

$$\mathbf{x} \precsim \mathbf{y} \Leftrightarrow \sum_i x_i \leq \sum_i y_i$$

We can now also state the Repugnant Conclusion more formally:[12]

The Repugnant Conclusion (Formal version). *For all* $y, z \in \mathbb{R}$*, where* $y > z > 0$*, and for any* $k \in \mathbb{N}$*, there is a* $n \in \mathbb{N}$ *such that* $(y)_k \prec (z)_n$.

[11] Although the assumption of completeness is standard in the population economics literature, some population ethicists have made attempts to avoid the Repugnant Conclusion by relaxing it. (See e.g. attempt by (Parfit 2016) and response by (Arrhenius 2016).) But, to keep things relatively manageable, we shall nevertheless in this paper assume completeness.

[12] This formalization is slightly different from that of (Blackorby, Bossert, and Donaldson 2005), who require that $n > k$. See (Spears and Budolfson 2021) for a discussion of heterogeneity in formalizations of the Repugnant Conclusion in the prior literature.

It is easy to verify that Total Utilitarianism implies the Repugnant Conclusion.[13] The Variable-Value views that we later discuss will be contrasted with both Average Utilitarianism (to be formally defined) and Critical-Level Generalized Utilitarianism (CLGU).[14] CLGU is a family of generalized total utilitarian views, which include for instance Critical-Level Utilitarianism, Prioritarianism, and Total Utilitarianism as special cases. The general view can be stated thus:

Critical-Level Generalized Utilitarianism (CLGU). *For any* $\mathbf{x},\mathbf{y} \in X$*:*

$$\mathbf{x} \precsim \mathbf{y} \Leftrightarrow \sum_i [g(x_i) - g(c)] \leq \sum_i [g(y_i) - g(c)]$$

where g is non-decreasing and non-convex, and $c \geq 0$ *is the critical level at wich adding a new life becomes a social improvement.*

Some views in this CLGU family avoid the standard formalization of the Repugnant Conclusion, at the cost of entailing another counterintuitive result, such as Arrhenius's (2000) family of sadistic conclusions (Franz and Spears 2020). CLGU is a significant family for our purposes because it is fully additively separable. That is, a CLGU better-than relation satisfies same-number, different-number, and existence independence axioms, always with a constant critical level (Blackorby, Bossert, and Donaldson 2005). So, because our calibration results depend on additive separability being violated, no CLGU view is subject to the calibration arguments of our paper.

Average Utilitarianism can now simply be stated as:

Average Utilitarianism (AU). *For any* $\mathbf{x},\mathbf{y} \in \mathbf{X}$*:*

$$x \precsim y \Leftrightarrow \bar{x} \leq \bar{y}$$

It can also be easily verified that Average Utilitarianism does not imply the Repugnant Conclusion. However, Average Utilitarianism is well-known to violate the Mere Addition principle, which we can now formally state as:

[13] Jacob M. Nebel (forthcoming) has recently formulated a version of totalism that avoids the Repugnant Conclusion, by including a lexical threshold in the conception of individual welfare. As our aim here is not to discuss the extent to which Total Utilitarianism implies the Repugnant Conclusion — but rather the extent to which Variable-Value views imply the Repugnant Conclusion — we will not discuss Nebel's or other totalist views that avoid repugnance.

[14] CLGU was introduced by (Blackorby and Donaldson 1984) and subsequently explored in depth by (Blackorby, Bossert, and Donaldson 2005) and (Bossert 2017).

> **Mere Addition** (Formal version). *For any* $\mathbf{x} \in X$*, and for any* $z \in \mathbb{R}$ *such that* $z > 0, \mathbf{x} \lesssim (\mathbf{x}, z)$.

Denying Mere Addition, for a complete ordering, is equivalent to entailing what we call the Anti-Natalist Conclusion:

> **Anti-Natalist Conclusion**. *There exists a* $z \in \mathbb{R}$*, where* $z > 0$*, and an* $\mathbf{x} \in X$ *such that* $(\mathbf{x}, z) \prec \mathbf{x}$.

In the remainder of this paper, we examine a novel weakening of Mere Addition that we argue is highly plausible. To state the principle, let us stipulate that there is well-being level beyond which lives at that level are excellent by the standards of 21st-century developed countries; and let $\mathbb{E} \subset \mathbb{R}$ be the set of well-being levels that are excellent by this same standard. For concreteness, let's set that level at 97.5th percentile of lifetime well-being in our current global population. To make things even more concrete, we shall occasionally assume that a typical professor in a developed country is at that level. And let $\mathbf{X}_R$ be the set of vectors that (we think) could realistically represent the lifetime well-being of the intertemporal world population. We can now finally state:

> **Weak Mere Addition**. *For any* $\mathbf{x} \in X_R$*, and for any* $y \in \mathbb{E}$*,* $\mathbf{x} \lesssim (\mathbf{x}, y)$.

While denying Mere Addition, for a complete ordering, "only" implies accepting the Anti-Natalist Conclusion, denying Weak Mere Addition in addition implies accepting a Strong Anti-Natalist Conclusion:

> **Strong Anti-Natalist Conclusion**. *There exists a well-being level* $y \in \mathbb{E}$ *and a population* $\mathbf{z} \in X_R$ *such that* $(\mathbf{z}, y) \prec \mathbf{z}$.

While we ourselves are sceptical of the Anti-Natalist Conclusion, we think that there is even stronger reason to reject the Strong Anti-Natalist Conclusion.

As we show in the next two sections, however, unless Variable-Value views imply the Strong Anti-Natalist Conclusion, when these views have been calibrated to plausible empirical assumptions about the size and welfare of the future world population, they must imply many instances of the Repugnant Conclusion. This follows from our novel application of a familiar logic in decision theory: calibration of variable-value objective functions to reveal tensions between intuitions for large-quantity decisions and intuitions for small-quantity decisions. The leading result in this literature is Rabin's (2000) celebrated argument about expected utility theory. Formally, we merely extend Rabin's argument about choice under risk to analogous functional forms in population ethics.

Rabin established that an expected utility maximizer can only be moderately risk averse when relatively small sums of money are involved — e.g. always turning down 50-50 gambles between losing $100 and winning $105 — if she is extremely risk averse when larger sums of money are involved — e.g. turning down 50-50 gambles between losing $2,000 and winning any (including infinite) amount of money. So, the lesson of Rabin's argument is that an expected utility maximiser is either surprisingly risk averse when stakes are large or surprisingly risk neutral when stakes are small.[15]

Our calibration result is that Variable-Value views are surprisingly anti-natalist when few extra lives are at stake (more specifically, they entail the Strong Anti-Natalist Condition) unless they are surprisingly totalist when more lives are at stake. (By "totalist" we mean making choices that should be found objectionable by those who reject Total Utilitarianism because of the Repugnant Conclusion. In other words, our result is that Variable-Value views can only accommodate an extremely weakened Mere Addition principle if they also imply countless instances of the Repugnant Conclusion.

Two prominent examples

This section turns to two prominent examples of Variable-Value population axiologies. Both of these are well-known in the literature to avoid traditional formalizations of the Repugnant Conclusion. We ask what these views recommend in repugnance-type tradeoffs between large and small populations, once calibrated to satisfy Weak Mere Addition and facts or predictions about the world population. Because they are algebraically tractable and well-studied, it is instructive to see why these particular versions of Variable-Value views imply instances of the repugnant conclusions, before considering (in the next section) a more general argument that applies to all Variable-Value views that violate the strong Mere Addition principle that Total Utilitarianism entails.

[15] (Nebel and Stefánsson 2020) apply a similar logic to inequality averse views about how to order populations of a fixed size, in particular, to Prioritarianism and Rank-Discounted Utilitarianism, and find that such views can only be moderately inequality averse when small differences in welfare are at stake — e.g. preferring that everyone is equally well off at level w to half the population being at level w– 0.9 while the other half is at level $w + 1$ — if they are extremely inequality averse when larger welfare differences are at stake.

Number-dampened generalized utilitarianism

The first view in the Variable-Value family that we shall consider can be stated as follows:

> **Number-Dampened Generalized Utilitarianism** (NDGU). *There is a* $\alpha \in (0,1)$ *such that for any* **x**,**y** ∈ X*:*
>
> $$\mathbf{x} \precsim \mathbf{y} \Leftrightarrow \bar{x}\, \mathcal{N}(\mathbf{x})^{\alpha} \leq \bar{y}\, \mathcal{N}(\mathbf{y})^{\alpha}$$

NDGU reduces the value of additions to the population as population size grows. One way to see this, and connect it to our formulation of the Repugnant Conclusion, is that $(y)_k \prec (z)_n$ if $\frac{y}{z} > \left(\frac{k}{n}\right)^{\alpha}$. From this we can see that as α approaches 0, the ratio between y and z needed for NDGU to imply that $(z)_n \precsim (y)_k$ becomes smaller; which is just another way to say that the closer α comes to 0, the closer NDGU gets to Average Utilitarianism, and the fewer instances of the Repugnant Conclusion it implies. However, because $\alpha > 0$, NDGU will imply some instances of the Repugnant Conclusion, in the sense that for *some intuitively large* difference between y and z, and for any k, there will be an n such that $(y)_k \prec (z)_n$. Moreover, this illustrates that while NDGU violates the strong version of Mere Addition that Total Utilitarianism implies (since $\alpha < 1$), it does imply some weak instances of Mere Addition. Our aim now is to explore what happens when we assume that it satisfies some particular (very plausible) instances of Weak Mere Addition.

This algebraic formulation applies a particular, one-parameter family of functional forms to the concave-transformation proposal of (Hurka 1983) and (Ng 1989). Here we expand our historical horizons: The relevant fact is that adding a well-off (by today's standards) professor lowers average lifetime well-being because there will be many future people who will be even better off. In making this assumption, we follow recent literature on the possible long-term human future, such as by (Greaves and MacAskill 2019) and (Ord 2020). If future lives are to be so many and so good, then α must be high if the addition of a well-off professor is not to be a worsening.

If there will be 2×10^{12} people overall, for example, and if the average lifetime well-being will be three times as high as that of the modern professor's, then that implies an α of at least 0.67. If the future is even better than that and the average person will be five times as well-off as our professor, then we have an α of 0.80 — greater because today's extra happy professor pulls the intertemporal average down by more.

These, too, imply repugnant-like quantitative consequences. For instance, $\alpha =$ 0.67 implies that a population of 31,695,627 people with lifetime well-being of 0.1 is better than a population of 1,000 people with lifetime well-being of 100. And $\alpha =$ 0.80 implies that a population of 5,639,614 people with lifetime well-being of 0.1 is better than a population of 1,000 people with lifetime well-being of 100.[16] Both of these better-than judgements (i.e., those entailed by $\alpha = 0.67$ and $\alpha = 0.80$) should, we contend, be found repugnant by those who oppose Total Utilitarianism due to the Repugnant Conclusion.

Rank-discounted generalized utilitarianism

The second Variable-Value view that we shall consider can be stated as follows:

> **Rank-Discounted Generalized Utilitarianism** (RDGU). *There is a $\beta \in (0,1)$ such that for any* $\mathbf{x},\mathbf{y} \in X$*:*
>
> $$\mathbf{x} \precsim \mathbf{y} \Leftrightarrow \sum_r \beta^r\, g\big(x_{[r]}\big) \leq \sum_r \beta^r\, g\big(y_{[r]}\big)$$
>
> *where g is increasing and weakly concave.*

A version of this view is defended by, for instance, (Asheim and Zuber 2014). It avoids the (universally quantified) Repugnant Conclusion because $\beta^1 + \beta^2 + \beta^3 \ldots$ is a convergent series, which ensures that the aggregated value a perfectly-equal population remains finite, no matter how large it becomes. Therefore, if k, in our formal statement of the Repugnant Conclusion, is sufficiently large, and if y is sufficiently larger than z, then there is *no* n such that, by RDGU, $(y)_k \prec (z)_n$. In other words, the (universally quantified) Repugnant Conclusion does not follow from RDGU. But, if β is sufficiently close to 1, then even large y could be part of an instance of the Repugnant Conclusion with a z that is small enough to capture the qualitative (intuitively repugnant) features of the Repugnant Conclusion.

RDGU does not satisfy the strong Mere Addition principle that Total Utilitarianism entails. This is because adding a life lowers the weights of any otherwise-existing higher-utility lives, which may reduce social welfare by more than the additional life increases it. However, RDGU must satisfy *some* Weak Mere Addition principle, since $\beta > 0$, which means that *some* mere additions are valuable. And, in fact, the closer β is to 1, the closer RDGU comes to implying the strong Mere Addition principle, in the sense of implying stronger instances of Mere Addition. We want to examine what RDGU implies if we assume that it satisfies particular (very plausible) instances of Weak Mere Addition.

To that end, we assume that a typical professor in the developed world is at the

[16] The reason why the number of people with 0.1 lifetime well-being that is needed to outweigh 1,000 people with lifetime well-being of 100 is smaller in the second case, is that as α becomes larger, the resulting axiology becomes closer to a totalist one.

97.5th percentile of lifetime well-being in our current world population,[17] so only 2.5 percent of people are better-off. That is around 182 million people. Under RDGU, adding such a professor reduces the weight on those 182 million very well-off people. The implication is that, if adding such a professor is not a worsening, β must be very close to 1. In particular, β must be greater than 0.99999995 if the 182 million people are all *no more than 1.0001 times as well-off* as our well-off professor, and even closer to one if the better-off people are even better-off than that (which is of course more realistic). β reaches 0.999999999 if the better-off people are a little more than six times as well-off as the professor, for example.[18]

Now consider what these high βs imply for repugnant-like tradeoffs. How many people each with a life at 0.01 would be needed to be better than a population with 100 people at 100? The answer ranges from 1,025,864 for $\beta = 0.99999995$ down to just above one million as β becomes closer to 1.[19,20] But those who believe that the Repugnant Conclusion must be avoided at all theoretical cost will presumably not be happy to have 1.026 million people at 0.01 instead of 100 people at 100. And yet, such is the consequence of choosing RDGU and maintaining that creating the happy professor is not a worsening given our actual world population.

A more general argument

The quantitative results of Section 3 depended upon two specific functional forms. Here we present a more general argument which applies to any variable-value goodness function of the form for any $\mathbf{u} \in \mathbf{X}$: $W(\mathbf{u}) = \bar{u} \times f(\mathcal{N}(\mathbf{u}))$, where f is positive, increasing, and concave.[21] Our arguments would extend readily to cases where $\bar{u}$ were replaced by a more general “equally-distributed equivalent” (Atkinson 1970), because the populations in the standard Repugnant Conclusion are perfectly equal. We however use the arithmetic mean for simplicity. NDGU is the special case where $f(\mathcal{N}) = \mathcal{N}^{\alpha}$; RDGU is the special case where $f(\mathcal{N}) = \frac{1-\beta^{\mathcal{N}}}{1-\beta}$, for perfectly-equal

[17] For simplicity we here ignore past and future generations, but the large number of future people who would, we assume, be better off than today’s happy professors only increases the force of this argument.

[18] Note that our argument can accommodate a positive critical level and a prioritarian transformation: simply subtract the critical level and/or do the prioritarian transformation before the rank-based weighting.

[19] The sum of a geometric series is $\frac{1-\beta^n}{1-\beta}$. Because well-being is constant in both populations, this requires solving for n such that $0.1 \times \frac{1-\beta^n}{1-\beta} > 100 \times \frac{1-\beta^{100}}{1-\beta}$.

[20] Here we follow the convention of normalizing the g measure around the 0 welfare level, that is, we assume that $g(0) = 0$.

[21] Note that such a view violates the strong Mere Addition principle entailed by Total Utilitarianism. Hence, the view examined in (Sider 1991) does not have this form. But, as mentioned in fn.7, there are strong normative reasons for excluding from consideration the view in (Sider 1991).

populations. Notice that the algebra of $\bar{u} \times f(\mathcal{N}(\mathbf{u}))$ resembles the algebra of (risk averse) expected utility — an affine probability times a concave von Neumann-Morgenstern transformation — which is why our argument follows from Rabin's (2000).

Figure 1 illustrates the argument. First, fix a lifetime well-being level for a weak mere addition assumption: the lowest lifetime well-being level that you are confident that, if added to our intertemporal population, would not decrease the goodness of the population. So far, we have been using the typical lifetime well-being of a developed-country professor, or more specifically the 97.5th percentile of the 2020 global socioeconomic distribution. But here, to get the strongest possible argument, we want the *lowest* lifetime well-being level for which mere addition reasoning can be safely applied. We expect that, for most readers, this is a level below that of most lives in 2020 developed countries. We denote this a well-being level of 1, where 0 is a neutral life (neither worth living nor worth not living).

Next fix a lifetime well-being level of what we expect the business-as-usual average lifetime well-being level to be for the overall intertemporal human population. By "business-as-usual" we mean the situation that will occur if the one life at well-being level 1 is *not* added. We label this average lifetime well-being level γ for "good" and assume that the long-run future of humanity is such that γ is greater than 1.

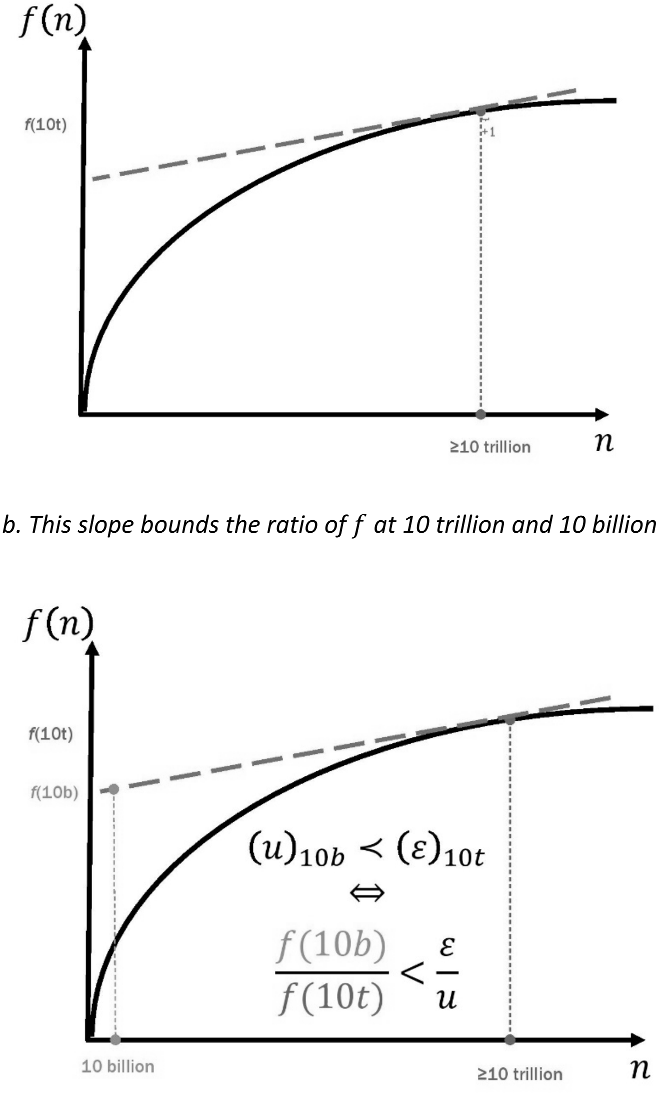
That it is not worse to add a life at 1, given the functional form of W, implies that:[22]

$$\frac{\gamma \times 10 \text{ trillion} + 1}{10 \text{ trillion} + 1} f(10 \text{ trillion } + 1) > \gamma f(10 \text{ trillion}).$$

This inequality bounds the slope of f at 10 trillion. By the concavity of f, the slope at 10 billion can be no less positive than the slope at 10 trillion. Extending this linearly towards zero provides an upper bound on f at 10 billion — but if concavity is steep the actual value may be well below this bound, resulting in even more repugnant-like implications.

[22] 10 trillion is an (Ord 2020)-type estimate of the plausible size of the intertemporal human population. 10 billion is a commonly-used size of the high-welfare population (A) in the Repugnant Conclusion literature.

Figure 1. A graphical representation of concave variable-value calibration



Panel a. Addition of a good life bounds the slope of f at 10 trillion

Panel b. This slope bounds the ratio of f at 10 trillion and 10 billion

For an explanation of the choice of population sizes, see fn. 22.

Now we are in a situation to draw quantitative repugnant conclusions. The precise numbers depend on γ.[23] We expect that readers will take γ to be large. Recall that γ is the ratio of long-term business-as-usual average lifetime well-being to the lowest lifetime well-being level such that we are confident that mere addition at that well-being level does not make the population worse. But our results are striking even if γ is not large.

- If $\gamma = 100$:
  - 10 trillion lives each at any positive well-being level worse than the threshold added life at 1 (or any other fixed, positive well-being level x) is better than
  - 10 billion lives, each 100 times as good as the threshold added life (or the other fixed, positive well-being level x).

- If $\gamma = 10$:
  - 10 trillion lives each at a well-being level $y > 0$ is better than
  - 10 billion lives, each 10 times as good as y.

- If $\gamma = 2$:
  - 10 trillion lives at any positive well-being level z is better than
  - 10 billion lives, each twice as good as z.

Clearly the implications for large γ constitute the very same repugnance that motivates some population ethicists to avoid Total Utilitarianism. To be sure, our argument uses an empirical premise about the size and well-being of the future population, formalized in the assumption that $\gamma > 1$. We believe that it is a strength of our argument that it speaks to a plausible calibration of the actual world. After all, that means that the implications we derive are not merely theoretical possibilities, but rather results that we would get if we were to apply the theories under examination in actual policy evaluation. But any reader uncomfortable with these empirical premises can read our argument as a *conditional* one, where the results are conditional on a plausible and relevant hypothetical future.[24]

Lesson and concluding remarks

Recall that the intuitive appeal of Variable-Values views was supposed to be that they could avoid the Repugnant Conclusion while satisfying at least some weak instance of the Mere Addition principle. We have now seen, however, that if these views satisfy what we take to be a very plausible, and certainly weak, instance of Mere Addition, and

[23] This dependence is because of the arithmetic of averages.

[24] See, e.g., (Ord 2020), (Greaves and MacAskill 2019).

if in addition we make plausible empirical assumptions about the intertemporal world population, then these Variable-Value views have implications that, we suggest, those who oppose Total Utilitarianism due to the Repugnant Conclusion will find repugnant.[25]

Why has the fact that Variable-Value views imply many instances of the Repugnant Conclusion been overlooked? We suggest that the reason is that standard formalizations of the Repugnant Conclusion use *universal* quantification ("For *any* perfectly equal population of very well-off people…"). But that quantification is not, we think, necessary to capture the intuition that there is something repugnant about views that suggest we choose a population consisting of lives that are barely worth living over a (smaller) population of excellent lives. That is, the fact that a view recommends we give up *many* populations of excellent lives for larger populations of lives that are barely worth living will strike those who worry about the Repugnant Conclusion as repugnant, even if the view in question does not make this suggestion for *all* populations. And as we have seen, Variable-Value population axiologies avoid *some* instances of such repugnant choices, but not other instances — just like any other population axiology (Spears and Budolfson 2021).

What should we conclude from our results? Most narrowly, a lesson of our results is that when calibrated to the real world — that is, the actual world population and what we think are plausible empirical assumptions about the future population — Variable-Value views substantially agree with Totalist views on how to rank policies that affect a relatively small number of people. Moreover, if we assume that policy choices typically affect only a relatively small number of people — that is, small in relation to the total intertemporal world population — then the implication is that these Variable-Value views and Totalist views typically recommend the exact same courses of action (especially if the menu of possible options is coarse). The only escape would be for these Variable-Value views to be strikingly anti-natalist, such that they do not even satisfy weak instances of Mere Addition that would involve the lives of well-off readers of this paper.

More broadly, these results teach us something about the effort to avoid the Repugnant Conclusion. Variable-Value axiologies are commonly taken to avoid the Repugnant Conclusion. However, these views cannot avoid supporting repugnant-type judgments; not only in theory, but also, as we have seen, when these views are calibrated to the real world. So, one lesson from our paper is that Variable-Value views have been excluded from the set of population axiologies understood to imply repugnance only because of how the Repugnant Conclusion is typically formalized

[25] In fact, given Tännsjö's principle of unrestricted instantiation (recall fn.9), these implications *must* be deemed repugnant if the Repugnant Conclusion is to be used as an argument against Total Utilitarianism.

and quantified — not because they would rank populations in a way that would seem satisfactory to those who find the Repugnant Conclusion repugnant.

Population ethicists have long understood that escaping undesirable or unintuitive implications is impossible. But this paper adds to a growing recent literature — including (Spears and Budolfson 2021) on additions to an unaffected population and (Arrhenius and Stefánsson 2018) on risky choice between uncertain populations — that finds repugnant conclusions even under approaches to population ethics commonly understood to avoid repugnance. Collectively, these results suggest that the effort to avoid the Repugnant Conclusion has, in some ways, hinged on questionable features of the formalization of repugnance (such as the features that exclude the cases documented in this paper); that some of this effort may therefore be misplaced; and that perhaps avoidance of the Repugnant Conclusion should not be a core goal of population ethics research.[26]

References

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

———. 2000. "An Impossibility Theorem for Welfarist Axiologies." Economics & Philosophy 16 (2): 247–66.

———. 2016. "Population Ethics and Different-Number-Based Imprecision." Theoria 82 (2): 166–81. https://doi.org/10.1111/theo.12094.

Arrhenius, Gustaf, and H Orri Stefánsson. 2018. "Population Ethics Under Risk." Working paper. IFFS.

Asheim, Geir B, and Stéphane Zuber. 2014. "Escaping the Repugnant Conclusion: Rank-Discounted Utilitarianism with Variable Population." Theoretical Economics 9 (3): 629–50.

Atkinson, Anthony B. 1970. "On the Measurement of Inequality." Journal of Economic Theory 2 (3): 244–63.

Blackorby, Charles, Walter Bossert, and David J Donaldson. 2005. Population Issues in Social Choice Theory, Welfare Economics, and Ethics. Cambridge University Press.

[26] The authors of (Zuber et al. n.d) — a recent statement of agreement by many authors from diverse perspectives — argue that avoiding the Repugnant Conclusion has been overemphasized by population ethics research.

Blackorby, Charles, and David Donaldson. 1984. "Social Criteria for Evaluating Population Change." Journal of Public Economics 25 (1-2): 13–33.

Bossert, Walter. 2017. "Anonymous Welfarism, Critical-Level Principles, and the Repugnant and Sadistic Conclusions." Working paper. University of Montréal.

Franz, Nathan, and Dean Spears. 2020. "Mere Addition Is Equivalent to Avoiding the Sadistic Conclusion in All Plausible Variable-Population Social Orderings." Economics Letters 196: 109547.

Greaves, Hilary, and William MacAskill. 2019. "The Case for Strong Longtermism." Working Paper. Global Priorities Institute.

Hurka, Thomas. 1983. "Value and Population Size." Ethics 93 (3): 496–507.

Nebel, Jacob M. forthcoming. "Totalism Without Repugnance." In Festschrift for Derek Parfit, edited by Tim Campbell, Jeff McMahan, and Ketan Ramakrishnan.

Nebel, Jacob M., and H. Orri Stefánsson. 2020. "Calibration Dilemmas in the Ethics of Distribution." Working paper. University of Southern California; Stockholm University.

Ng, Yew-Kwang. 1989. "What Should We Do about Future Generations?: Impossibility of Parfit's Theory x." Economics & Philosophy 5 (2): 235–53.

Ord, Toby. 2020. The Precipice: Existential Risk and the Future of Humanity. Hachette Books.

Parfit, Derek. 1984. Reasons and Persons. Oxford.

———. 2016. "Can We Avoid the Repugnant Conclusion?" Theoria 82 (2): 110–27.

Pivato, Marcus. 2020. "Rank-Additive Population Ethics." Economic Theory 69 (4): 861–918.

Rabin, Matthew. 2000. "Risk Aversion and Expected-Utility Theory: A Calibration Theorem." Econometrica 68 (5): 1281–92.

Sider, Theodore R. 1991. "Might Theory x Be a Theory of Diminishing Marginal Value?" Analysis 51 (4): 265–71.

Spears, Dean, and Mark Budolfson. 2021. "Repugnant Conclusions." Working paper; prior version is IZA Discussion Paper 12668.

Tännsjö, Torbjörn. 2020. "Why Derek Parfit Had Reasons to Accept the Repugnant Conclusion." Utilitas, 1–11.

Zuber, Stéphane, Dean Spears, Johan Gustafsson, Mark Budolfson, and others. n.d. "What Should We Agree on about the Repugnant Conclusion?" Working paper. UT Austin.

Joe Roussos[1]

Awareness Growth and Belief Revision[2]

The problem of awareness growth, also known as the problem of new hypotheses, is a persistent challenge to Bayesian theories of rational belief and decision making. Cases of awareness growth include coming to consider a completely new possibility (called expansion), or coming to consider finer distinctions through the introduction of a new partition (called refinement). Recent work has centred on Reverse Bayesianism, a proposal for rational awareness growth due to Karni and Vierø. This essay develops a "Reserve Bayesian" position and defends it against two challenges. The first, due to Anna Mahtani, says that Reverse Bayesian approaches yield the wrong result in cases where the growth of awareness constitutes an expansion relative to one partition, but a refinement relative to a different partition. The second, due to Steele and Stefánsson, says that Reverse Bayesian approaches cannot deal with new propositions that are evidentially relevant to old propositions. I argue that these challenges confuse questions of belief revision with questions of awareness change. Mahtani's cases reveal that the change of awareness itself requires a model which specifies how propositions in the agent's old algebra are identified with propositions in the new algebra. I introduce a lattice-theoretic model for this purpose, which resolves Mahtani's problem cases and some of Steele and Stefánsson's cases. Applying my model of awareness change,

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

[2] Financial support by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences) is gratefully acknowledged.

then Reverse Bayesianism, and then a generalised belief revision procedure, resolves Steele and Stefánsson's remaining cases. In demonstrating this, I introduce a simple and general model of belief revision in the face of new information about previously unknown propositions.

Introduction

In ordinary life and in science, we regularly confront new possibilities. When I moved to Stockholm, I learned about that distinctive Swedish item of cutlery, the smörkniv—a smooth wooden "knife" used exclusively for butter. Climate scientists in the twentieth century developed and explored new theories and mechanisms, such as the runaway greenhouse effect, to understand and explain observed climate phenomena. In different ways, these each involved the formation of new beliefs. The fact that Swedes use elegant juniper wood implements for spreading butter is a prosaic proposition that I had simply never encountered before. In the climate case, the scientists involved learned entirely new concepts and theories.

It is a strange failing of our formal models of belief that they have little to say about this kind of learning. Consider "Bayesian" models of belief, those that represent beliefs with probabilities and insist that learning is accomplished by conditionalization. Bayesianism s a rich and successful theory (if so broad a church can be called a "theory"), both in philosophy and statistics. But in Bayesian models, all resolutions of uncertainty take place by updating pre-existing beliefs. Agents must have priors for propositions to learn about them at later stages. In this way, Bayesianism leaves no room for agents to learn about genuinely new states of affairs and has no guidance for real agents when they undergo such changes of awareness. And yet, such new possibilities arise in some of the most important challenges facing us today, from climate change to existential risk. What to believe, and how to act, in the face of this kind of uncertainty is a matter of great import.

I will set out a model of growing awareness that specifies how an agent's probabilistic beliefs ought to be extended to a new space of possibilities and provides constraints for the formation of new beliefs about the new possibilities. My focus will be entirely on degree of belief, leaving desire and decision aside.

The problem I am interested in here has been discussed previously under the description "the problem of new hypotheses" in the philosophy of science (it is linked to the "problem of old evidence" raised by Glymour (1980). As the name implies, the focus there was on hypotheses and their confirmation, which limited the scope of the discussion somewhat. Previous work in that context includes that by Shimony (1970), Eels (1985), Earman (1992), and more recently Wenmackers and Romeijn (2016)). My problem is the probabilistic companion to a much discussed issue in logic and computer science, concerning the logic of unawareness and qualitative belief revision following awareness growth. Schipper (2015) is a thorough review. As my focus is on probabilistic belief, my attention will be on a more recent literature in decision theory and epistemology, where the problem has come to be known as "awareness growth".

Our starting point is a probabilistic model of rational belief. A probabilistic model of belief requires at least two components: an algebra of propositions and a probability function. It is usual to talk about the probability function as representing the agent's beliefs, or degrees of belief. Awareness, as I will think of it, is a necessary condition for taking an attitude toward a proposition.[3] As the probability function represents the attitude (belief), it is a natural suggestion to use the algebra itself to model the agent's awareness. At first pass we will say that a proposition within the algebra is one that the agent is aware of, and any propositions not in the algebra are those the agent is unaware of.[4] The algebra also encodes logical relationships between propositions. Changes to awareness will be thought of as, in the first instance, changes to the algebra. (This sets up a helpful parallel between the problem of probabilistic belief extension and revision in the face of new possibilities, and familiar techniques of qualitative belief revision pioneered by Levi (1977, 1980), Alchourrón, Gärdenfors, and Makinson (1985).) I will idealise and assume that the agent is logically omniscient: their degrees of belief reflect all the deductive logical relationships between propositions of which they are aware.[5]

As the probability function is defined on the algebra, this modelling choice means that in our model agents have no attitudes to propositions they are unaware of. This, I claim, is what we want. I'm interested in what we might call *true unawareness*, and true awareness growth: situations in which agents confront genuinely new possibilities. The common cases of this include learning a new concept, considering a new scientific theory, learning things about a city you've never visited before, and many others.

This is a narrower definition of awareness growth than others in the literature have used. Bradley, Steele and Stefánsson consider cases in which agents have forgotten or failed to consider a proposition, or in which they deliberately exclude a proposition from consideration, as cases of unawareness and treat remembering or

[3] It may also be a sufficient condition, but I think there is reason to doubt this. Consider a comparativist framework in which comparative partial beliefs are primary and probabilistic credences mere representations of them. If incomplete partial beliefs are rationally permitted then it seems possible to me that an agent might be aware of a proposition without their partial belief relation including it. Even though it would be in the domain of an imprecise probabilistic representor of that agent, we might still want to say that the agent does not truly have an attitude to that proposition. See footnote 4 for a related comment.

[4] This is no constraint on extending the model to include desire and decision, so long as we adopt a propositional decision theory such as that developed by Richard Jeffrey (1983).

[5] This is an important idealisation. I am not attempting to solve all the problems of Bayesianism at once; rather I am isolating awareness growth as a particular challenge, and showing how it can be resolved. With this idealisation in place, we can make use of a comment by Isaac Levi (1991) to illuminate the nature of awareness. Just as Levi points out that a logically closed belief set contains the sentences an agent is *committed to believe*, the awareness state contains propositions that the agent is committed to take some credal attitude to.

considering as awareness growth. My restriction to true awareness growth is less severe than it may seem. While agents who have, for example, forgotten something are not truly unaware of the relevant propositions, they may be fruitfully *modelled as if they were* for the purposes of an analysis. We might use a model of unawareness to study the rational constraints on forgetting, and a model of awareness growth to study remembering. Nevertheless, we can recognise a significant conceptual difference between an agent who is truly unaware of a proposition—such as a scientist in the moments just prior to first learning of a new theory—and an agent who is temporarily excluding something from consideration.

A complication in this discussion is that the literature operates in terms of toy models involving small sets of prosaic propositions, even when discussing true unawareness. It is hard to shake the intuition that surely the agents know something about the ″new″ propositions in these stories—*obviously they know about buses going to town*, or *surely they had heard of French films before*. This is a problem with the examples, and with the intuitions attached to them, when it comes to studying true unawareness. This is further compounded by modelling cases where the agent does know about the bus going to town, *but has forgotten it,* as unawareness. I shall pick my way through this thicket by considering only true unawareness, and forcefully rejecting those hangover intuitions in toy examples.

The general problem that I consider is how an agent's beliefs should rationally change when they become aware new propositions. The proposal I defend is as follows. An agent's initial awareness state is represented by the Boolean algebra on which their probabilities are defined. After their awareness grows, their beliefs are again represented by probabilities defined on an algebra. The old algebra is related to the new by a kind of mapping called a lattice embedding: a one-to-one homomorphism which preserves logical conjunctions and disjunctions, but not negations. The agent's initial probabilities are ″extended″ to the new algebra, a process which determines what the old belief state has to say about the wider set of possibilities the agent now confronts. This process of extension takes place without considering any new evidence that the agent learns during the experience that brings about her awareness growth. After her initial probabilities have been extended to the new algebra, they can be updated in a belief revision process to reflect any information she has learned about the new possibilities, or the relations between new and old possibilities.

Reverse Bayesianism

The approach just outlined is a development of a recent line of investigation, beginning with economists Karni and Vierø (2013, 2015, 2017), and developed in philosophy by Richard Bradley (2017) . My development is motivated in part by recent criticisms of this approach by Steele and Stefánsson and Mahtani. I will therefore briefly introduce Karni and Vierø and Bradley's original proposals.

Karni and Vierø's is a complex, choice-based treatment of growing awareness, but I will interpret the epistemic portion of their proposal into a propositional framework as follows.

> **Reverse Bayesianism.** Suppose that A and B are maximally-specific propositions the agent was previously aware of, P represents the agent's prior beliefs, and P^+ represents their extension after a growth of awareness. For any such A, B, where $P(A) > 0$ and $P(B) > 0$, a rational agent will have:
>
> $$\frac{P(A)}{P(B)} = \frac{P^+(A)}{P^+(B)}$$

Reverse Bayesianism (RB) is presented (e.g., by Vallinder (2018) and Steele and Stefánsson) as a form of conservative belief revision: in the face of new information, the agent responds appropriately to the new information while preserving as much as possible from their prior belief state. In this case, what is preserved is the ratio of probabilities of propositions they were previously aware of.

In some ways RB is a weak constraint. It tells you only how to constrain the ratios of probabilities of propositions you were previously aware of. It says nothing about the probabilities of new propositions. Another way of stating this is that the permissible posterior belief state, according to RB, is highly imprecise: it is a set of probability functions, each of which obeys RB, but which differ on the new propositions. This is sensible: an agent who was previously unaware of certain possibilities has no real way of constraining their attitudes toward those possibilities.[6]

Here is an example of the principle in action.

> *Weather*. Naledi is considering tomorrow's weather. Being South African, she is aware of three possible kinds of weather: rain, clouds, sun. But having just moved to Sweden, she becomes aware of a fourth kind of weather: snow. Her awareness

[6] At the moment, this may sound more polemical to precise Bayesian ears than I intend it to be. What I mean is: there's nothing in the *prior* that will help fix the probabilities of the new possibilities. If you're committed to determining unique credences using the principle of indifference, have at it. As will become clear later, there is no conflict with how I interpret Reverse Bayesianism.

> state grows from the three propositions represented by {RAIN, CLOUDS, SUN} to the four propositions {RAIN, CLOUDS, SUN, SNOW}[7]

Let's assume that for Naledi these are mutually exclusive, maximally-specific, propositions. RB demands that, for example, $P^{+}(\text{RAIN})/P^{+}(\text{CLOUDS}) = P(\text{RAIN})/P(\text{CLOUDS})$, but says nothing about $P^{+}(\text{SNOW})$.

Richard Bradley's proposal is similar. Bradley argues that the conservatism we see in three Bayesian belief revision rules (Bayes, Jeffrey and Adams updating) involves the rigidity of conditional beliefs. So, he concludes, extending probabilities to a wider algebra s1hould also preserve conditional probabilities. He provides a condition which does this and which vindicates certain intuitions about awareness growth in simple cases. The condition is that "the agent's new conditional probabilities, given the old domain, for any members of the old domain should equal her old unconditional probabilities for these members" (Bradley 2017, 258). Or, to use terminology Bradley introduces, the new belief states must be *rigid extensions* of the old.

Here's a fuller definition of Bradley's proposal:

> **[Rigid Extension.]** Let $\Omega = \langle \mathcal{X}, \vDash \rangle$ be a Boolean algebra of propositions of which the agent is initially aware and let $\bigvee\mathcal{X}$ be its top element. Let P be a probability function defined on Ω. Let $\mathcal{E}$ be some set of propositions not contained in $\mathcal{X}$. Let $\mathcal{Y}$ be the closure of $\mathcal{X} \cup \mathcal{E}$ under the Boolean operations and $\Xi = \langle \mathcal{Y}, \vDash \rangle$, be a Boolean algebra of prospects based on $\mathcal{Y}$. Note that $\bigvee\mathcal{X}$ belongs to $\mathcal{Y}$. Then, for any P, a corresponding P^{+} on Ξ is called a *rigid extension* of P to Ξ iff, for all $X \in \mathcal{X}$,
>
> $$P^{+}(X|\bigvee\mathcal{X}) = P(X)$$

Bradley takes it to be a norm of rationality that our beliefs are rigidly extended when our awareness grows.

Consider Weather again. Let $\mathcal{X} =$ {RAIN, CLOUDS, SUN} be Naledi's initial awareness state. Then if P^{+} is a rigid extension of her prior P , it will have

[7] Although awareness states are represented by Boolean algebras, when discussing toy examples I'll often refer to the partition of maximally specific propositions that the agent is aware of *as* their awareness states, metonymically.

$P^+(\text{RAIN}|\bigvee\mathcal{X}) = P(\text{RAIN})$, but again there are no constraints on $P^+(\text{SNOW})$.[8]

While not equivalent to Reverse Bayesianism, Bradley's rigid extension (RE) principle does imply it, on one reading.[9] This has led some authors, such as Katie Steele, Orri Stefánsson, and Anna Mahtani, to *call* Bradley's position Reverse Bayesianism. In any case, counterexamples to RB will invalidate RE.

RE is, roughly, the proposal that I think is correct for awareness growth. But as it stands, its formulation is unclear and underspecified, as the following section will demonstrate.

The key insight in the RE proposal is that awareness growth is (almost) the inverse problem to learning. In a learning problem, an agent excludes possibilities by learning which element of a partition is true. By conditioning on the learned proposition, the agent restricts the support of their credence function to a subset of the initial algebra: the algebra "below" their evidence—the logically strongest proposition that they have credence 1 in.[10]

In an awareness growth case, the agent includes new possibilities and shifts to a wider algebra. Their initial beliefs, defined on the smaller initial algebra, provide limited constraints on this new set of possibilities. In the absence of new *evidence* (i.e., when all that has been introduced is new possibilities), the principle of rationality guiding their extension is this: if the agent later discovers that the new possibilities are not the case, and restricts their beliefs to the set of propositions corresponding to their prior awareness state, they should recover their initial beliefs. (This is "reverse Bayesianism" in the sense that extension should allow some hypothetical subsequent Bayesian learning to recover your current state.) Why? Because they have no new information which could motivate a change in the contents of their beliefs. All they have gained is an awareness of possibilities, and not any evidence regarding those possibilities, nor any evidence about the old possibilities.

[8] Everything I argue for in this paper is compatible with an imprecise probability representation of belief. Indeed, that is the account I favour and it is the context in which Bradley introduces his Rigid Extension principle. I work in terms of precise priors here for simplicity only.

[9] Here's a misleading "proof". Consider the ratio $P^+(A|\bigvee\mathcal{X})/P^+(B|\bigvee\mathcal{X})$, for $A, B \in \mathcal{X}$ with $P(A) > 0, P(B) > 0$. Note that $A \vDash \bigvee\mathcal{X}$ and $B \vDash \bigvee\mathcal{X}$, Thus $P^+(A \wedge \bigvee\mathcal{X}) = P^+(A)$ and $P^+(A \wedge \bigvee\mathcal{X}) = P^+(A)$. So $P^+(A|\bigvee\mathcal{X})/P^+(B|\bigvee\mathcal{X}) = P^+(A)/P^+(B) = P(A)/P(B)$. My later arguments will show why this reasoning is faulty, but I think that it is assumed by the authors I am responding to.

[10] This is one way of thinking of it, in any case—see for example Miklos Redei's work on conditional expectation. If we want to use the algebra to model awareness, we may need to revisit this way of thinking. An agent who learns by conditioning doesn't lose their awareness of the possibilities that they can now exclude. Similarly, they're aware of all of the entailments of their evidence but these sit "above" the top element of the sub-algebra I describe. These issues don't matter for my main argument here, but they are worth investigating.

Awareness Growth

In the literature on awareness growth there is much informal talk about "growing" the agent's algebra, and about the new algebra "containing the old propositions." In this section I will highlight various problems with this informal discussion, and how it has been treated in formal models.

The background analogy underlying such talk seems to be something like a set of everyday objects. You start off with two apples, and describe them as a set of two apples. You place a third next to them, and note that now you have a set of three. Physically, the original two are still there, and there's one new apple next to them. Similarly, we might say that the set of two apples "grew" and the set of three apples contains the original two.

For cases like Weather, there's some initial plausibility to this: it seems natural to say that Naledi's partition of precipitation propositions went from {RAIN, CLOUDS, SUN} to {RAIN, CLOUDS, SUN, SNOW}—the old propositions, RAIN and so on, are right there next to the new one, SNOW. This kind of case is called an *expansion* in the literature: a partition is expanded by adding a new proposition.

Another archetypical case is distinguished in the literature: *refinement*. Here, a new partition is introduced and the agent comes to make finer distinctions than they did before.

> *Weather 2*. Naledi now has a four element precipitation partition, as before. But she realises that she needs to consider temperature too. She distinguishes two temperatures: hot and cold. Her awareness state grows from the four propositions {RAIN, CLOUDS, SUN, SNOW}, to the eight propositions represented by {RAIN, CLOUDS, SUN, SNOW}∧{HOT, COLD}.[11]

Naledi's awareness has been refined because, where she previously distinguished one possibility RAIN, she now distinguishes two, RAIN∧HOT and RAIN∧COLD.

But now we can see that the simple notion of propositional preservation across awareness changes will not do. For one thing, cases like Weather 2 immediately illustrate the importance of certain modelling choices that are usually unimportant.

Suppose that propositions are thought of as sets of possible worlds. In Weather 2, we initially need four worlds, which we can label in terms of the propositions Naledi is initially aware of: {RAIN, CLOUDS, SUN, CLOUDS}. When the temperature partition is added, this completely changes the world-structure needed to model Naledi's awareness. Whereas the proposition RAIN was previously a singleton, consisting of a

[11] This notation is to be read as follows: for two sets $\mathcal{X}, \mathcal{Y}, \mathcal{X} \wedge \mathcal{Y} = \{x \wedge y : x \in \mathcal{X}, y \in \mathcal{Y}\}$.

RAIN-world, it becomes a set of two worlds: {RAIN∧HOT, RAIN∧COLD}. The RAIN proposition has different content after the awareness growth. Talking about it as being "the same", and even applying the label "HOT", may be misleading (this has been pointed out by Steele and Stefánsson forthcoming).

Alternatively, we might take propositions to be fundamental and use "worlds" as merely a colourful description for maximally specific propositions. On this way of thinking, we can say "nothing has changed about the propositions RAIN—it remains unchanged by the awareness growth—all that has happened is that we've introduced a new partition and so there are more maximally specific propositions than there used to be, as these are just elements of the joint partition."

But even here there are complications. Consider the proposition RAIN, and its negation ¬RAIN. Before her awareness grew for the first time, Naledi took ¬RAIN to be CLOUDS∨ SUN. But after her awareness grows, ¬RAIN is CLOUDS∨SUN∨ SNOW. We're in a strange position: we can say that the RAIN proposition is "still there" in the new algebra, but its negation has changed. On one way of specifying the proposition, as ¬RAIN, we can say it is "still there"—RAIN of course has a negation. But it is equivalent to different propositions in the old algebra and the new algebra.

We may want to say that the issue here has to do with negation (indeed, this is what I will say later). But we also have to worry about the interaction between how we label and structure propositions in our model, and facts about the agent. In some contexts, it is perfectly legitimate to relabel propositions for convenience, and to talk about whichever partitions are useful for our purposes. Is that the case here? If so, life is more complicated. Suppose that instead of speaking in terms of the partition {RAIN, CLOUDS, SUN} we shift our attention to {DRY, ¬DRY}, where "DRY" is a label used to track all of the events that don't involve rain. So initially ¬DRY is a relabelling of RAIN, while DRY is a equivalent to CLOUDS∨SUN. Now when Naledi becomes aware of the possibility of snow, she distinguishes it from rain and so SNOW ends up in DRY alongside CLOUDS and SUN. So, after Naledi's awareness grows, it is the "plain" proposition DRY which has changed, while its negation ¬DRY remains the same.

How are we to build these new algebras? Richard Bradley (2017, 258–59) makes a simple proposal for how to construct Ξ from Ω. He supposes that the agent becomes aware of a set of propositions $\mathcal{U}$, with $U \notin \mathcal{X}$, for all $U \in \mathcal{U}$. We start by forming Y, the closure of $\mathcal{U} \cup \mathcal{X}$ under the Boolean operations. Then $\Xi = \langle \mathcal{Y}, \vDash \rangle$ is a Boolean algebra, which Bradley calls the *extension* of Ω by $\mathcal{U}$. Note that $\bigvee\mathcal{X} \in \mathcal{Y}$, and in general $\bigvee\mathcal{X} \neq \bigvee\mathcal{Y}$.

(For ease, I'll use a closure operator to refer to algebras in terms of sets that generate them: for a set X, $\mathrm{cl}(\mathcal{X})$ is the closure of $\mathcal{X}$ under the Boolean operations $\wedge$,$\vee$, $\neg$. So the closure of $\{A\}$ is the simplest algebra $\{\bot, A, \neg A, \top\}$.)

Bradley's extension proposal is underspecified. Consider Weather again. Naledi becomes aware of $\mathcal{U} = \{SNOW\}$. Bradley's proposal advises us to take the union of this set with $\mathcal{X}$, the set containing the propositions already in her algebra. In this case, $\mathcal{X} = \mathrm{cl}(\{RAIN, CLOUDS, SUN, SNOW\} \wedge \{HOT, COLD\})$. But the set theoretic operation of union is not going to do the job alone. What we want is for SNOW to go into the precipitation partition. It wouldn't do to have SNOW end up in the temperature partition, nor for it to be completely independent so that we get a $\{SNOW, \neg SNOW\}$ partition.

This question of specifying a procedure for constructing the new algebra is slightly tangential to the main discussion here, and so I will set it aside for the moment and assume that we're doing the construction by hand so that we get the right algebra for each toy case we look at.

Note that these difficulties have nothing to do with belief, or belief revision—there's no mention of P in the few paragraphs above. They concern how to describe changes of awareness at the level of the propositions and algebra involved, and in particular how to identify propositions across such changes. These issues do, however, generate problems for belief revision procedures such as Reverse Bayesianism because the authors who proposed them have not yet resolved these prior questions satisfactorily.

Mahtani's "splitting" proposition cases

In a recent paper, Anna Mahtani (2020) presents a problem for RB and RE that turns on these concerns. She introduces two awareness growth cases which, looked at one way, involve expansion, while looked at another way, involve refinement. As a result, RB appears to generate two conflicting demands for preserving probability ratios, in a way that disallows assigning any credence to the new possibilities. Here is the first case.

> *The Other Tenant*. Suppose that you are staying at Bob's flat which he shares with his landlord. You know that Bob is a tenant, and that there is only one landlord, and that this landlord also lives in the flat. In the morning you hear singing coming from the shower room, and you try to work out from the sounds who the singer could be. At this point you have two relevant propositions that you consider possible...with LANDLORD standing for the possibility that the landlord is the singer, and BOB standing for the possibility that Bob is the singer.
>
> Because you know that Bob is a tenant in the flat, you also have a credence in the proposition (TENANT) that the singer is a tenant. Your credence in TENANT is the same as your credence in BOB, for given your state of awareness these two

> propositions are equivalent. Let us suppose, just for simplicity, that your credence in LANDLORD is 0.5 and your credence in TENANT (and so of course in BOB) is 0.5.
>
> Now let's suppose that the possibility suddenly occurs to you that there might be another tenant living in the same flat, and that perhaps that is the person singing in the shower. Let's assume that no other possibilities occur to you—e.g. it does not occur to you that it might be a visitor singing in the shower, or just a recording, or anything like that.

Mahtani says: "we have a refinement relative to the possibilities LANDLORD and TENANT, but an expansion relative to the possibilities LANDLORD and BOB." (p.7)

As "TENANT and BOB are possibilities that you were aware of before the awareness growth," Mahtani considers applying RB to the TENANT-BOB pair in which case they have the same credence and OTHER gets none. She then considers applying RB to the LANDLORD-BOB and LANDLORD-TENANT pairs, but then again RB will leave no credence for OTHER. Finding this result unreasonable, Mahtani takes it as evidence that Reverse Bayesianism cannot be the correct norm for awareness growth.[12]

Before the awareness growth, the alternative to LANDLORD is a single proposition which I will label TENANT/BOB. After the awareness growth there are two alternatives to LANDLORD: BOB and OTHER. Reverse Bayesianism says that the probability ratios between familiar propositions must be preserved. In so doing, it assumes that all of the old propositions are in the new algebra, and that there's no ambiguity about which of the propositions in the new algebra they are. But in this example, it appears that a single proposition from the old algebra (TENANT/BOB) has been replaced by two propositions in the new algebra (TENANT and BOB), in a hierarchical relationship: TENANT has two sub-possibilities, BOB and OTHER. This allows one to find multiple pairs of propositions to feed into the RB recipe, leading to the incorrect prescription that OTHER must receive no credence.

This isn't quite how Mahtani put it, of course. At times, she talks about TENANT and BOB as two names for the same proposition (before the awareness growth). But at other times, she says that you were aware of two propositions before the change: TENANT and BOB. It is just that, "given your state of awareness these two propositions are equivalent." While I find it difficult to state clearly, I think I understand what this means. Before the awareness change, you can distinguish between the

[12] Mahtani says that Reverse Bayesianism "effectively rules out awareness growth" in this case. I dispute this language use: it seems perfectly possible that one becomes aware of a possibility and simultaneously learns that it is not the case. Awareness is modelled by the proposition's presence in the algebra, not by the probability it is assigned.

claim that Bob is singing and the claim that the tenant is singing (their utterances use different words, they have different senses), and perhaps you recognise that for some speakers they won't be substitutable. Their "equivalence" means that, given what you're aware of, you take them to have the same referent and your attitude to them is the same. I think one could argue that they are the same proposition, but I don't want to go too deeply into a discussion of the nature of propositions.

I also do not wish to disparage the import of Mahtani's challenge to Reverse Bayesianism. I agree that we should expect any account of awareness growth to be able to handle this sort of case, and that RB and RE cannot do so at present. I diagnose RB's problem with these cases differently, however. I think RB is underspecified because it doesn't properly deal with how the algebra itself has changed. In its current form, it admits of at least two specifications. On Mahtani's specification, when an old proposition "splits" into two in the new algebra, the RB rule must be applied to the two split propositions (which means they get equal credence) and to all the pairs these split propositions form with other old propositions. This gets the wrong prescription. On another specification, we must choose one, and only one, of TENANT and BOB to be "the same as" TENANT/BOB and apply RB accordingly. This won't generate the incorrect prescription, but the problem is how to specify which new proposition to identify with the old.

Mahtani considers a proposal by Steele and Stefánsson for restricting RB to pairs of propositions X, Y such that the new possibilities are evidentially irrelevant to the matter of X versus Y. The introduction of OTHER is irrelevant to LANDLORD versus BOB, on their proposal. But as OTHER is relevant to TENANT, the pairs involving TENANT aren't subject to RB. This gets the intuitively correct result, that the relative probability of LANDLORD and BOB stays fixed, and OTHER is assigned any value of probability you like within that constraint.[13]

But Mahtani introduces another example that they cannot handle, a case with the same structure but a different intuition about which ratio should be preserved. In The Other Tails, an agent considers a coin and wonders whether it will land HEADS or TAILS. They then think about what image is on the tails side of the coin. Initially they think all coins show a lion, so that TAILS and LION are equivalent (as TENANT and BOB were taken to be above). Later, the agent comes to consider the possibility that the tails side has an image of Stonehenge. So in the new awareness state, TAILS has two sub-possibilities, LION and STONEHENGE. The revised Reverse Bayesian prescription is that you must keep your credence in HEADS and LION in the same proportion as before. But here, intuitively, we want the proba-

[13] I'll come to Steele and Stefánsson in more detail later on, but for the moment want to continue drawing out the implications of Mahtani's examples.

bility ratio of HEADS-TAILS to be preserved—after all, what the image might be is irrelevant to the probability of a coin toss. But HEADS-TAILS is like LANDLORD-TENANT, and in The Other Tenant we concluded that we wanted the LANDLORD-BOB ratio to be preserved. We need an explanation of how our intuitions are settling on which pair's ratio should be preserved, and our rule for belief revision needs to track that explanation.

Again I think that what this shows is that Reverse Bayesianism, and the account that underlies it, is under-specified. There is something right about it—in each case, there is a probability ratio we want preserved. But it doesn't tell us which one that is, nor is it accompanied by a compelling story of why it should be that ratio. I will argue that a fuller analysis of the awareness change, as distinct from the accompanying belief revision, will help to clarify matters and allow us to preserve what is good about RB.

A model of awareness growth (but not belief revision)

What this shows us is that talking about awareness growth, and specifically about changes to the algebra, is difficult and confusing when done in this informal manner. We need a precise way of defining what it means for a proposition from the old algebra to be "in the new algebra", or put another way, of defining what it means for a proposition from the old algebra to be identical with a proposition in the new algebra.

In order to do that, I will now introduce some mathematical machinery. The agent's awareness state is modelled by a Boolean algebra, which philosophers are used to thinking of as a kind of field of sets. I will instead consider them algebraically, by which I mean that I will be thinking of a Boolean algebra as a complemented distributive lattice of propositions, ordered by an implication relation.

Here is a brief introduction to lattice theory terminology. Considered algebraically, a lattice is a mathematical structure $\langle \mathcal{X}, \wedge, \vee \rangle$, consisting of a set $\mathcal{X}$ and two operations, called meet ($\wedge$) and join ($\vee$). Meet and join are associative and commutative, each is idempotent, and they obey an absorption law.A distributive lattice is one where the meet and join operations distribute over one another. A bounded lattice has distinguished elements, $\perp$ and $\top$, here called bottom and top respectively, such that $X \wedge \perp = \perp$, and $X \vee \top = \top$ for all $X \in \mathcal{X}$. A complemented distributive lattice is a bounded lattice such that, for each $X \in \mathcal{X}$, there is a unique element of $\mathcal{X}$, denoted $\neg X$, such that $X \wedge \neg X = \perp$ and $X \vee \neg X = \top$. Note that, while the distinguished elements are typically denoted $\top$ and $\perp$, in the context of multiple algebras it is useful to make the algebra explicit and thus to denote them $\bigvee \mathcal{X}$ and $\bigwedge \mathcal{X}$ respectively.

In our case, $\mathcal{X}$ will be a set of propositions. The notation above is deliberately suggestive: meet and join are equivalent to the logical operations of conjunction and disjunction. In some contexts I will talk in terms of an implication relation, $\vDash$, defined by $X \vDash Y$ iff $X \wedge Y = X$ iff $X \vee Y = Y$. $\vDash$ is also called the ″order″ for the lattice.

Lattices can be visualised in Hasse diagrams. A Hasse diagram represents each element of the lattice, and draws a line that goes upward from X to Y whenever $X \vDash Y$. There is no interpretation of lines crossing, all the matters is the start- and end-point of any line. Three simple lattices are shown in Figure 1. Note some differences between the three examples. In 1(a), $P \vee Q = \top$, so if we take this to be a *complemented* lattice, $\neg P = Q$. In 1(b), by contrast, $P \vee Q \neq \top$, and instead $\neg P = Q \vee R$. In 1(c), I've shown a case where P and Q are independent propositions, each with their own complement labelled in the usual way.

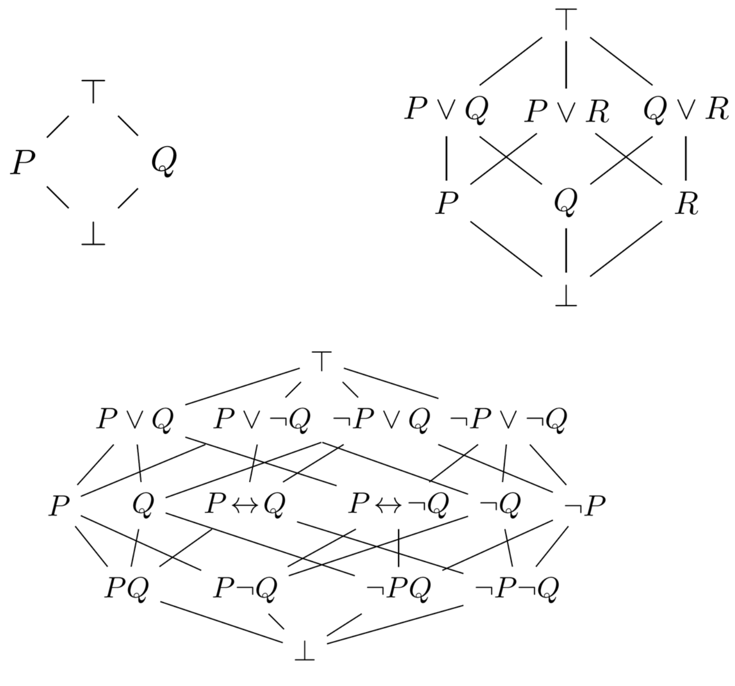


Figure 1. Hasse diagrams showing three simple lattices. From left to right: (a) Two non-trivial elements in a single, two-element partition. (b) Three atoms in a single, three-element partition. (c) The Boolean algebra $cl(\{P, Q\})$.

So, Boolean algebras are particularly rich examples of lattices. As is well known, this richness is crucial to defining probabilities. In the probabilistic belief models I am interested in, Boolean algebras represent awareness states: each element of the algebra is a proposition of which the agent is aware. In cases of awareness change, both the initial and final awareness states will be represented by Boolean algebras. The question this section addresses is: what is the relation between the initial and final algebras? This is the reason that I am interested in lattice theory, which mostly considers sets equipped with less structure than full Boolean algebras: because awareness *changes* only partially preserve the Boolean algebra structure.

Recall the Weather example above, where after Naledi's awareness change, the negated proposition ¬ RAIN had changed. This is because in some kinds of awareness growth (often called *expansion* cases), the agent comes to realise that what they previously took to be the total set of possibilities is incomplete. Naledi realises that {RAIN, CLOUDS, SUN} isn't a complete set of precipitation states—and so it isn't certain that RAIN∨ CLOUDS∨SUN. While RAIN∨CLOUDS∨SUN was the top element of the algebra representing Naledi's initial awareness state, it is not the top element of the algebra that represents her new awareness state.

We want to relate the old algebra to the new, in a way that tracks our intuitions about which propositions are "still there" while also tracking this change of complementation structure. For this we need a mapping between the two. As we're considering awareness growth, where new possibilities are added, what we need is a mapping called a *lattice embedding*.

> **Lattice embedding.** A map $h: \mathcal{L} \rightarrow \mathcal{K}$, between two lattices $\langle \mathcal{L}, \vee, \wedge \rangle$ and $\langle \mathcal{K}, \vee, \wedge \rangle$, is a lattice embedding iff it is one-to-one lattice homomorphism. That is, a one-to-one map that is meet- and join-preserving: $\forall X, Y \in \mathcal{L}, h(X \vee Y) = h(X) \vee h(Y), h(X \wedge Y) = h(X) \wedge h(Y)$.

A lattice embedding maps each proposition from the old algebra to a proposition in the new algebra, and which preserves the lattice operations, meet and join.[14] Above I've made their respective meet and join operations explicit, and put a bar on the operations in the codomain algebra. The fact that that it is one-to-one means that the image $h(\mathcal{L})$ is a sublattice of $\mathcal{K}$ which is isomorphic to $\mathcal{L}$. Note that a lattice embedding does *not* preserve complementation, i.e., there is no guarantee that $h(\neg X) = \neg h(X)$.

In Weather, Naledi's initial precipitation partition is {RAIN, CLOUDS, SUN},

[14] Relating to the order-theoretic definition of lattices: defined this way, lattice homomorphisms are order-preserving (Davey and Priestley 2002Proposition 2.19, p.44)

and her final precipitation partition is {Rain, Clouds, Sun, Snow}. (In order to make the point I want to here, I've changed the capitalisation of the final versions.) The natural embedding maps h(RAIN)=Rain, h(CLOUDS)=Clouds and h(SUN)=Sun. The fact that *h* is a lattice homomorphism guarantees that h(RAIN ∨ CLOUDS) = Rain ∨ Clouds and h(RAIN ∧ SUN) = Rain ∧ Sun. In this latter case, the conjunction of RAIN and SUN is a contradiction, RAIN ∧ SUN = ⊥, and our mapping gets us this too as, in the new algebra, Rain ∧ Sun = ⊥. However, in the initial algebra ¬ RAIN = CLOUDS ∨ SUN, whereas in the new algebra ¬ Rain = Clouds ∨ Sun ∨ Snow. So h(¬ RAIN) = h(CLOUDS ∨ SUN) = h(CLOUDS) ∨ h(SUN) = Clouds ∨ Sun ≠ ¬ h(RAIN). As advertised, the lattice embedding does not preserve complementation which tracks the result of our informal model and our intuition about what has changed for Naledi.

We've now arrived at our recipe for modelling awareness growth.

> **Modelling recipe.** An agent's awareness state is modelled by a Boolean algebra $\Omega = \langle \mathcal{X}, \vee, \wedge \rangle$. After their awareness grows, they have a new awareness state: the Boolean algebra Ξ. We relate the old algebra to the new by embedding Ω into Ξ. The one-to-one association of propositions in Ω with propositions in Ξ ensures that the old propositions are "in" the new algebra: there is an embedding *h* such that, for each $X \in \Omega$, $h(X) = x \in \Xi$. (Ξ must of course also contain the new propositions that are the content of the awareness change.)

The recipe does not tell us, at this stage, which embedding we need.

As this model will do significant work in criticising previous authors writing on awareness growth, I want emphasise why we should represent things this way. First, when comparing algebraic structures and their elements, the natural notion of "identity" is isomorphism. The lattice embedding is an isomorphism between the old algebra and its image in the new algebra. It is the right kind of isomorphism, because it preserves meet and join but not complementation. Second, I have highlighted difficulties with the informal, intuitive notion of identity between propositions in different algebras, so *some* new account is needed. I claim that mine has various benefits. It gives a clear answer to what it means to identify a proposition from the old algebra with a proposition in the new algebra. It also explains what is puzzling about Mahtani's cases, and provides a resolution.

Mahtani's cases, again

Let's start by thinking about Mahtani's statement that there are three propositions in your initial awareness state in The Other Tenant: LANDLORD, TENANT, and

BOB. In a lattice theoretic presentation we can't represent BOB and TENANT as two different propositions, at best they're two labels for the same proposition. That's because propositions are elements of the set $\mathcal{X}$ that is the basis of the algebra Ω, and sets contain only one copy of each item. Or, from another direction: because Ω is a Boolean algebra, each element must have a *unique* complement—if TENANT and BOB have the same complement (LANDLORD) then they must be identical.

I don't mean to hide behind the formalism here; this isn't meant to be a "computer says no" kind of response. The whole theory of probability is built on Boolean algebras, and these just aren't the kind of structures that can represent the kind of distinct-but-equivalent propositions that Mahtani gestures at. I don't think that is a problem either. As I will show, we can capture the sense in which the original "TENANT" proposition "splits" when your awareness grows.

The definition of identity comes directly from the lattice homomorphism. Consider again Ω and Ξ, with h a lattice embedding between them. A proposition $Y \in \Xi$ is "the same as" a proposition $X \in \Omega$ just in case h maps X to Y: $h(X) = Y$. As h is one-to-one, every proposition in Ω is associated with one in Ξ, and as it is a homomorphism, the disjunction and conjunction relations between propositions are preserved.

What is puzzling about Mahtani's cases is that they offer us two choices for how to embed the original algebra into the new algebra. But *this problem has nothing to do with Reverse Bayesianism.*

In Figure 2(a), I've shown the initial algebra Ω with the two possibilities you start off being aware of: L for LANDLORD and TB for TENANT/BOB. In this model, there is simply no way to represent Mahtani's description of the initial awareness state as containing a proposition BOB and another TENANT, which coincide in some sense. The best we can do is to recognise the two ways of labelling the one proposition, which I've done with the composite label TB. Figure 2(b) and (c) show the new algebra Ξ, in which there's a composite possibility, TENANT (t), with two sub-possibilities, BOB (b) and OTHER (o). I've used lower-case labels for these propositions in Ξ, because I want to insist that they're mathematically different entities, which will come to be associated with the elements of Ω only via an embedding.

Here is one possible embedding $h: \Omega \to \Xi$, $h(L) = l, h(TB) = b$. Under this embedding, the old possibility TENANT/BOB is mapped to the new possibility BOB. We can work out the rest from the fact that h is a lattice homomorphism. So $h(\bot) = h(L \wedge TB) = h(L) \wedge h(TB) = l \wedge b = \bot'$. Importantly, $h(\top) = h(L \vee TB) = h(L) \vee h(TB) = l \vee b$, which is not the top element of Ξ! That is, of course, what we want: you previously thought that the Landlord and Bob were the only possibilities for the singer, but then you learned that there is another possibility. This also changes the complementation structure of the algebra: now you realise that, if it is not the Landlord singing, then it might be Bob or the Other Tenant. Previously you thought that if it were not the Landlord, it must be Bob. The image of Ω in Ξ, under this embedding, is shown in Figure 2(b) in bold. This embedding makes The Other Tenant an example of expansion.

Another embedding h' maps TENANT/BOB to the new possibility TENANT. It is shown in Figure 2(c) in bold. On this embedding, The Other Tenant is a case of refinement: where you previously thought in terms of "the tenant" you now recognise two finer distinctions within this proposition.

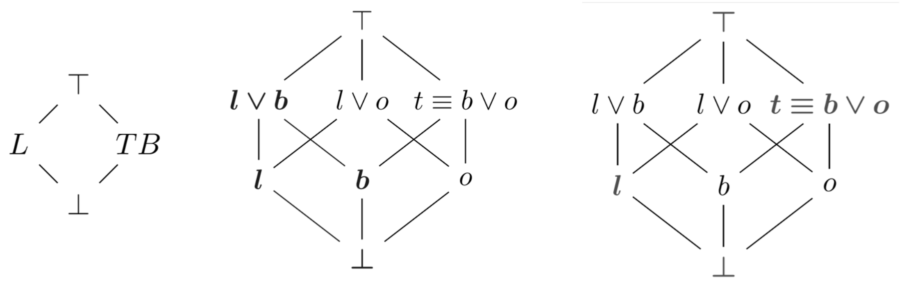


Figure 2. The Other Tenant. From left to right: (a) The old algebra. (b) The new algebra, with the old embedded into the bold portion. (c) The new algebra, with a different, bold, embedding. On Mahtani's analysis, (b) is the preferred embedding for The Other Tenant.

Although this section is just about awareness changes and "growing an algebra", it will help see the purpose of all this is we take a quick detour into belief revision. (Mahtani, after all, is arguing against Reverse Bayesianism.) First, note that the RB principle itself needs to be adjusted. It is framed in terms of "the same proposition" appearing in the old and new algebra, which we now see is just a shorthand; convenient but potentially misleading. Here is a revised principle, tracking the relevant lattice embedding:

> **Reverse Bayesianism (new).** Let Ω be an agent's initial algebra, and Ξ their final algebra, related by a lattice embedding $h: \Omega \to \Xi$. Let P be the agent's prior credence function on Ω, and let P^+ be a probability function on Ξ. For any $A, B \in \Omega$, where $P(A) > 0$ and $P(B) > 0$, P^+ is a rational extension of P to Ξ iff:

$$\frac{P(A)}{P(B)} = \frac{P^{+}(h(A))}{P^{+}(h(B))}$$

If we apply Reverse Bayesianism (new) to the two embeddings I introduced above for The Other Tenant, we get different results. On h (left) we get:

$$\frac{P(L)}{P(TB)} = \frac{P^{+}(h(L))}{P^{+}(h(TB))} = \frac{P^{+}(l)}{P^{+}(b)}$$

Whereas on h' (right) we get:

$$\frac{P(L)}{P(TB)} = \frac{P^{+}(h'(L))}{P^{+}(h'(TB))} = \frac{P^{+}(l)}{P^{+}(b \vee o)}$$

Notice that we don't ever get the situation Mahtani discusses: where we are forced to assign TENANT and BOB the same probability by RB. That's because embeddings are one-to-one mappings, so there is no embedding which will map TENANT/BOB to TENANT and to BOB. The sense in which TENANT/BOB "is" the proposition TENANT *and* "is" the proposition BOB is just that there exists an embedding on which it is mapped to each of them. So our revised RB principle won't problematically assign the new possibility zero credence, in the way that Mahtani highlighted.

We can now evaluate these two embedding options. (We would like a systematic way of doing this, but for now I'll keep it intuitive and informal.) The h embedding requires that there is no change in the relative probabilities assigned to the singer being the Landlord or Bob. So, whatever credence is assigned to the possibility that there is another tenant who is the singer, it needs to get its probability mass equally from that previously assigned to LANDLORD and TENANT/BOB. Recalling that Mahtani has the priors set up with $P(TB) = P(L) = 0.5$, this means that the extended probability LANDLORD and BOB must be identical: $P^{+}(l) = P^{+}(b) = 0.5 - k, P^{+}(o) = 2k$.

The h' embedding, on the other hand, requires that the extended probability for LANDLORD and TENANT be identical. TENANT has two sub-possibilities, BOB and OTHER. So where you previously assigned equal credence to the singer being the Landlord or the tenant (who you took to be Bob), now that you are aware of the possibility of another tenant, you assign equal credence to the singer being the landlord or *either* tenant: $P^{+}(l) = P^{+}(t) = 0.5$ as before. There are no constraints on the credence assigned to b and o so long as $P^{+}(b) + P^{+}(o) = 0.5$.

The latter assignment seems, to me, to be the less sensible. Mahtani agrees: ″given that there might be two tenants, it is natural to suppose that your credence in TENANT should increase relative to LANDLORD″ (p.9). Incidentally, it is also the result given by Steele and Stefánsson's restriction of RB.

In The Other Tails, we face a similar choice. Recall that in this example, you previously had two labels, TAILS/LION, for one of your propositions. When your awareness grows, you end up with a proposition TAILS that has two sub-possibilities, LION and STONEHENGE. We can therefore embed your old algebra into the new one using a (suitably defined) h embedding, shown in bold in Figure 3(b), or with an h' embedding shown in bold in Figure 3(c). In this case, however, our intuitions are that h' is the more sensible embedding: as I said above, the number of potential images on the tails side doesn't change the probability of the coin toss result.

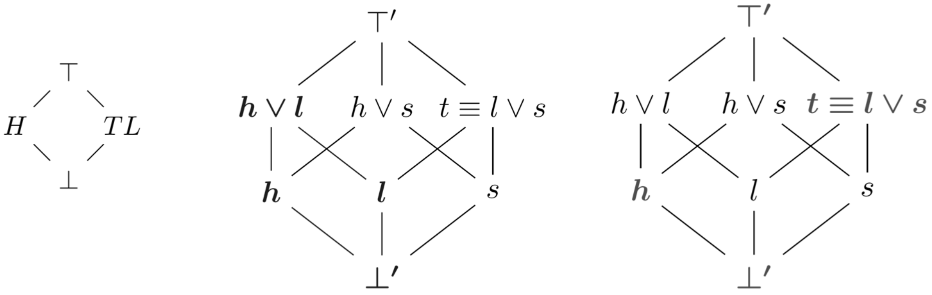


Figure 3. The Other Tails. From left to right: (a) The old algebra. (b) The new algebra, with the old embedded into the bold portion. (c) The new algebra, with a different bolded embedding. On Mahtani's analysis, (c) is the preferred embedding for The Other Tails.

Choosing an embedding

While the structure that I have provided above clarifies the problem of awareness growth significantly, we would still like guidance on which embedding is the right one for a particular example. This would be valuable philosophically, and is of practical necessity for any who would wish to translate my eventual proposal into algorithms for machine learning purposes. Unfortunately I am not sure this kind of exercise can be solved in perfect generality. What I can offer is some guidance, based on an examination of the cases above and the reasoning that seems to underlie the intuitions about the correct embeddings.

I begin with The Other Coin. This example has a particular feature that the others do not: the agent's subjective probabilities for the coin toss are derived from the chances we commonly take to hold for coin tosses. These chances plausibly derive from the symmetry of the coin itself: its two sides and constant weight distribution. They therefore ″attach″ to the {HEADS, TAILS} way of specifying the possibilities. When the agent extends their beliefs to the new algebra, they ought to do so in a way which respects the chance structure that determined their original credences. So this is a case in which we are choosing an embedding—which is a way of specifying how the agent's *awareness* has grown—based on belief considerations. In particular we look at the reasons that their beliefs are structured as they are, and choose an embedding which allows us to carry these reasons over to the new algebra.

Now let us consider The Other Tenant. Here, Mahtani tells us that the agent starts by assigning equal probability to LANDLORD and TENANT/BOB. If we take the lesson from the coin case to apply here, we should look for the reasons that underlie this assignment of credence and see how they apply to our choice of embedding. Often such assignments of equal credence are due to reasoning according to a ″principle of indifference″, on which one assigns equal probability to each outcome. Can this allow us to distinguish between the {LANDLORD, BOB} partition and the {LANDLORD, TENANT} partition? Perhaps. Applications of the principle of indifference require specifications of ″outcomes″ which are considered equivalent. On the {LANDLORD, BOB} way of thinking, the reasoning is presumably that there are two people in the house, that there is no reason to suppose one of them is more likely to be showering at the moment, and therefore to assign them equal probability. On the {LANDLORD, TENANT} way of thinking, the reasoning would have to involve the two roles in the legal agreement governing Bob's occupation of this apartment. It seems much less plausible that the agent's initial reasoning attached itself to the legal roles of ″landlord″ and ″tenant″ than to the two individuals qua people. Therefore, when we consider which embedding to choose on the basis of the reasons that fixed the agent's prior beliefs, the embedding which identifies propositions on the basis of personhood is preferred to the embedding which identifies propositions on the basis of legal role.

This, then, is my proposal about embedding selection: we choose the embedding which best preserves the reasoning underlying the agent's initial credal assignments.

Belief Revision

In this section I will show how to handle belief revision in cases where the agent learns new information as part of their awareness growth experience. In so doing, I show how the model above, along with the multi-stage analysis of awareness growth experiences that it fits naturally with, delivers the right answer to a set of cases that Steele and Stefánsson (forthcoming) have recently introduced as challenged for the Rigid Extension principle.

Steele and Stefánsson's evidential relevance cases

Steele and Stefánsson claim that RE is violated by ″examples where awareness grows since an agent becomes aware of a proposition that she takes to be evidentially relevant, intuitively speaking, to the comparison of propositions of which she was already aware″ (p.55). They offer a few examples of this sort, but I will focus on the one I take to be the most challenging.

> *Relativity*. Nineteenth century physicists were unaware of the Special Theory of Relativity [SR]. That is, not only did they not take the theory to be true; they had not even entertained the theory. We can suppose, however, that they had entertained various propositions for which the theory was regarded evidentially relevant, once Einstein brought the theory to their attention. In particular, they did (rightly) take the theory to be evidentially relevant to various propositions about the speed of light, such as whether the speed of light would always be measured at 300,000 km/s independently of how fast the investigator is moving or whether the measured speed would differ, depending on how fast the investigator is moving. But then the awareness and subsequent acceptance of [SR] changed their relative confidence in such propositions. (p. 55–56)

As Steele and Stefánsson interpret the example, the agent understands that SR is evidentially relevant to the proposition ″light's speed is constant″, LC. The thought is that, after their awareness grows to include SR, we don't want $P^+(LC)/P^+(\neg LC) = P(LC)/P(\neg LC)$, but rather we expect that $P^+(LC)/P^+(\neg LC) > P(LC)/P(\neg LC)$, as SR is a theory on which light's speed is absolute, and so any allocation of probability mass to SR will make that LC more likely than it was before. They conclude ″that we should not impose Reverse Bayesianism as a general constraint on how a rational agent can revise her credences when her awareness grows″ (p.57).

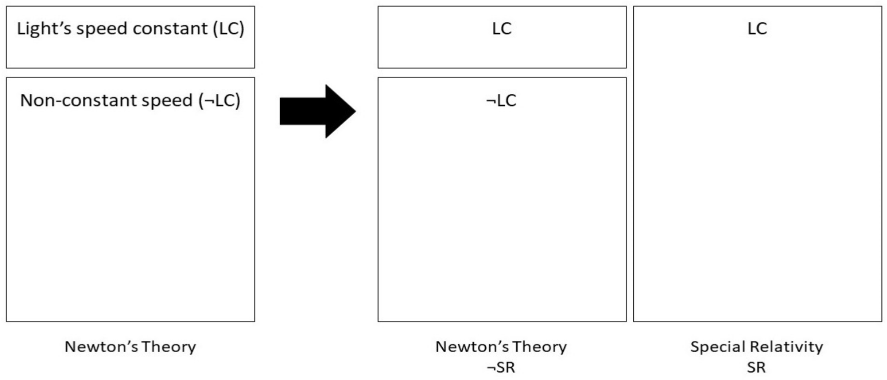


Figure 4. Recreation of Steele and Stefánsson's diagram showing the awareness growth in Relativity, as they see it.

Let us note the nature of the example. Steele and Stefánsson speak of the new proposition (SR) being "evidentially relevant" to an old one (LC), but in this case the link is stronger. The constancy of the speed of light is a postulate of Special Relativity, and so SR logically entails LC. (This is relevant to the lattice structure of the new algebra. As SR entails LC there will not be a SR∧ ¬LC proposition in the algebra. In the standard model of probabilistic beliefs, we process logical entailment in a different way from evidential relations. The former are modelled in the structure of the underlying algebra, while the latter are modelled in the agent's conditional probabilities.)

I will argue that there are two problems with Steele and Stefánsson's analysis of the example. Firstly, they fail to properly appreciate the requirements for identifying old propositions with new, as outlined in section 3. Secondly, and independently, their analysis only holds if one insists that RE cover all the changes that need to take place following awareness growth. In other words, if one assumes that these cases must be analysed in one go, rather than in stages as I described above, and if one assumes that RE is a proposal for such a single-stage analysis.

Let us begin with the familiar. The way in which Steele and Stefánsson take Relativity to be a counter-example to Bradley's Rigid Extension reveals why we need the embedding framework I introduced above. Recall that Rigid Extension requires that the agent's new conditional probabilities, given the top element of the old algebra, for any members of the old algebra, should equal her old unconditional probabilities for these members. Here is an updated version of this principle, reflecting the lattice embedding:

Rigid Extension (new). Let $\Omega = \langle \mathcal{X}, \vDash \rangle$ be the agent's initial awareness state, and P be their prior credence function. Let $\Xi = \langle \mathcal{Y}, \vDash \rangle$ be the new algebra after their awareness grows, and P^+ be their extended credences on Ξ. Then P^+ is called a *rigid extension* of P iff, for all $X \in \mathcal{X}$,

$$P^+(h(X)|h(\vee\mathcal{X})) = P(X)$$

Steele and Stefánsson in drawing out some differences between a version of RE and Reverse Bayesianism, write: "[Figure 4] makes vivid that [Rigid Extension] does not generally entail Reverse Bayesianism, at least for cases of awareness growth by expansion. [RE] requires that the probabilities for old propositions—[LC] and [¬LC]—*conditional on 'Newton's theory'*, remain constant when awareness grows. One can see just by looking at the figure that it does not follow that the ratio of the absolute probabilities for [LC] and [¬LC] remain constant when awareness grows" (p.64).

On my embedding view, Steele and Stefánsson's application of RE to the Relativity case is incoherent. In order to associate the old propositions with the new, we need to specify an embedding to take us from the initial awareness state to the new one. But their description of the problem corresponds to no possible embedding. To see why, consider Figure 5: 5(a) shows the initial awareness state, in which there is only one theory and two possibilities concerning the speed of light (lc and $\neg lc$). As in their diagram, Newton's theory is the only option and therefore is equivalent to the tautology. In Figure 5(b) we have the new algebra, in which there are two theories (SR for Special Relativity, and ¬ SR for Newton's theory), and two propositions concerning light (here, LC and ¬LC). The task before us is to embed the algebra shown on the left into the algebra on the right.

In the quote above, Steele and Stefánsson implicitly use an embedding on which the top element of the initial algebra ($\top$) is mapped to "Newton's theory" (¬SR) in the refined algebra *and* which maps "light is constant" (lc) to the full LC proposition in the new algebra. No homomorphism can accomplish this. Any embedding h on which $lc \mapsto$ LC will preserve the will map $\top \mapsto \top'$, because it needs to preserve the join structure:

$$h(l \vee \neg l) = h(l) \vee h(\neg l) = \text{LC} \vee \neg\text{LC} = \top'$$

It is possible to map $\top \mapsto$ ¬SR, but then "light is constant" must be mapped to the proposition beneath ¬ SR: $lc \mapsto \neg\,\text{SR} \wedge \text{LC}$. Indeed, this seems sensible: before becoming aware of Special Relativity the scientist took Newton's theory to be the only option, but after their awareness grows they see it is merely one of two possible accounts: $\top \mapsto \neg\mathrm{SR}$. But on this way of thinking, you have to give up the identity between the old "light is constant" proposition and the new, higher up, proposition denoted LC. You can't have it both ways.

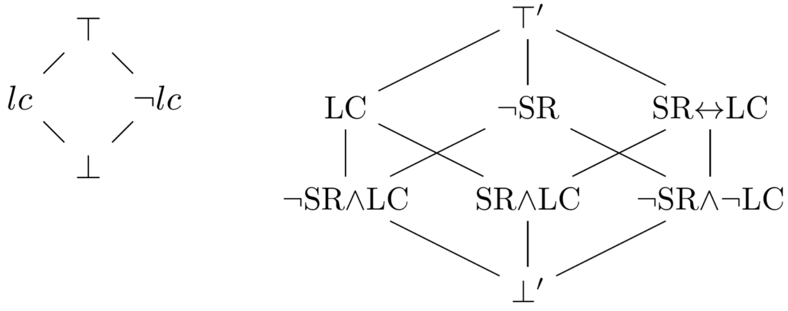


Figure 5. Hasse diagrams for Relativity. From left to right: (a) the initial algebra, and (b) the refined algebra after awareness growth.

Now on either embedding we can apply Rigid Extension. Due to the structure of the algebra, any positive probability assigned to SR (which is equivalent to SR∧LC in the diagram) will raise the probability of LC. But the Rigid Extension principle simply doesn't have anything to say about what probability one assigns to new propositions such as SR. So there is no conflict. One might worry that this case doesn't illustrate the point quite in the way Steele and Stefánsson want, because LC is a postulate of Special Relativity, and so the evidential relevance is handled entirely by the structure of the new algebra. What it means for a logically omniscient agent to "become aware" of SR and LC is to fit them into an algebra such as that shown in Figure 4(b), and then to map the old algebra into it with an embedding. As SR entails LC, any suitable algebra will have the SR proposition sitting below it.

But when the relevance of a new proposition to an old is merely evidential, rather than logical, this won't occur. It is instructive to end with such a case, to see how we handle agents who learn *probabilistic* information about the relations between new propositions and old. Consider the following case.

> *Movies.* Suppose you are deciding whether to see a movie at your local cinema. You know that on the day in question, the cinema only shows "international" (non-English) movies. You realise that both the movie's language and genre will affect your viewing experience. The possible languages you consider are French and German and the genres you consider are thriller and comedy. But then you realise that, due to your poor French and German skills, your enjoyment of the

> movie will also depend on the level of difficulty of the language. Since you know the owner of the cinema to be simple-minded, you are, after this realisation, much more confident that the movie will have low-level language than high-level language. Moreover, since you associate low-level language with thrillers, this makes you more confident than you were before that the movie on offer is a thriller as opposed to a comedy. (Steele and Stefánsson forthcoming, 58)

Steele and Stefánsson write: "The important feature of the above example is that the original awareness context is partitioned according to some property (the language level) that is taken to be evidentially relevant to the comparison of some pair of inconsistent basic propositions—that the movie is a thriller and that it is a comedy—in the old awareness context" (Steele and Stefánsson forthcoming, 58). RE will say that, in such cases, the ratio of probability between Thriller and Comedy ought to remain constant. But, argue Steele and Stefánsson, this should not be—instead we should end up with the credal assignment represented schematically in Figure 6.

I begin again with a comment on the nature of the case. This is a case of forgotten or neglected information, in which the agent not only has previous acquaintance with the "new" propositions, but also to evidential relations between "new" and old propositions. Steele and Stefánsson propose to model it *as if* it were awareness growth, which I agree is possible. But to do so we must treat the "new" information as if it were really new. We therefore proceed as follows: we identify the initial and final algebras, we find an embedding to map the old algebra into the new, we extend the agent's probabilities by RE, and then we apply the new information that is "learned" (in this case, remembered) in a belief revision stage.

Here the "learned" information has two parts. Firstly, "you associate low-level language with thrillers", which we can model as coming to have $Q(\text{Thriller} \mid \text{Low})=1$. (Note that in this case we assume this is a contingent empirical fact.) Secondly, "you know the owner of the cinema to be simple-minded, you are, after this realisation, much more confident that the movie will have low-level language than high-level language": $Q(\text{Low}) > Q(\text{High})$. Each is a piece of information that "comes to you" in your remembering.

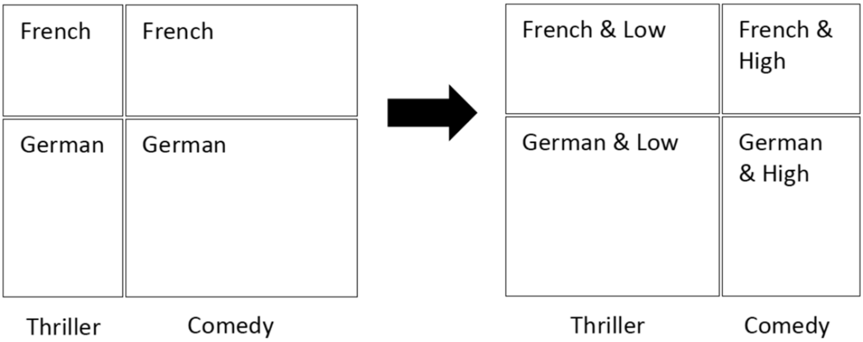


Figure 6. Steele and Stefánsson's diagram showing the awareness growth in Movies, as they see it.

I do not deny that Figure 6 shows the intuitively correct result given all the information presented in Movies. However, I propose that RE should be regarded as only a partial specification of the change to the agent's belief state. Specifically, it deals with the change due *only* to the change of awareness. Put another way, it is a rule for *extending* belief following awareness growth, not for revising it in light of new evidence. Belief revision will be handled separately.

On this view, RE says nothing about how you should change your belief upon learning that a new possibility is evidentially relevant to an old possibility. So it shouldn't be expected to hold *after* new things are learned. It is neither a ban on future learning, nor a one-stop-shop for belief revision in the face of growing awareness.

Semi-formally (i.e., without employing my embedding model) here is how I would apply the rule in this case. The awareness growth takes the agent from an awareness state {French, German}∧{Thriller, Comedy} to final awareness state {French, German}∧{Thriller, Comedy}∧{Low, High}. It is at this stage that RE (or indeed RB) applies. It tells us how to extend the agent's probabilities to the new algebra, and entails that we shouldn't change the ratio of probabilities of known propositions. I.e., the probability ratio of Thriller to Comedy should remain constant. But that isn't the end of the story. The agent also comes to have Q(Thriller | Low)=1 and Q (Low) > Q (High). Once these changes are incorporated, it is clear that $Q(\text{Thriller}) > Q(\text{Comedy})$, as expected.

There are three stages to the analysis. The first stage involves apprehending the new propositions and recognising their logical relations to known propositions. It is modelled by forming the relevant algebras and choosing an embedding which associates the known propositions with propositions in the new algebra. With this done, we move to the second stage and extend the prior attitudes to the new awareness state in line with the demands of rationality. (I agree with Bradley that the right demand is RE.) Having done so, we can update the agent's beliefs in light of the new information learned during the experience in the third stag. The separation is conceptual rather than temporal: the aim is to focus first on the purely awareness-related aspects of the experience and then to turn to the attitude of belief. The experience that brings about these three steps might (and often will) be a single, unitary experience.

The purpose of the separation is analytical clarity, allowing us to distinguish between the agent's priors on the old algebra (P), their extension to the new algebra (P^+), and the posterior after learning (Q). This distinction is necessary: the agent can't be updating P with the new information, as it isn't defined on the right algebra.

Steele and Stefánsson are aware of this option. They reject (a two-stage variant of) it on grounds of ad hocness:

> It might be argued that our examples are not illustrative of a simple learning event (a simple growth in awareness); rather, our examples illustrate and should be expressed formally as complex learning experiences, where first there is a growth in awareness, and then there is a further learning event that may be represented, say, as a Jeffrey-style or Adams-style learning event. In this way, one could argue that the awareness-growth aspect of the learning event always satisfies Reverse Bayesianism (the new propositions are in the first instance evidentially irrelevant to the comparison of the old basic propositions). Subsequently, however, there may be a revision of probabilities over some partition of the possibility space, resulting in more dramatic changes to the ratios of probabilities for the old basic propositions. The reason we reject this way of conceiving of the learning events described by our examples is that the two-part structure is ultimately unmotivated. The second learning stage is an odd, spontaneous learning event that would be hard to rationalise. Hence, this would again seem to us to be an artificial and ad hoc way to save Reverse Bayesianism. (p.59)

"Spontaneity" seems like the wrong protest. There is nothing "spontaneous" about the learning event—it is brought about by the experience of, for example, becoming aware of the new scientific theory. It also seems fairly plausible that in many cases one *does* become aware of something before learning about its evidential relations to other matters.[15] Whether it is "hard to rationalise" will depend on our account of belief revision, which I turn to in the next subsection.

This is one place where my focus on *true* unawareness is important. In their more prosaic examples like Movies, Steele and Stefánsson draw on background knowledge of evidential relations between new propositions and old. As an agent comes to consider a possibility (which they were previously aware of but neglected), they employ this background knowledge. But these "new" propositions aren't new in the sense I'm interested in, they were merely forgotten or neglected until this point. If we want to model such cases as if they were awareness changes, we need to treat the evidential relations as something learned. Not because they are truly learned anew, but because we're using a model developed for a different kind of case (true awareness growth) to model prosaic cases about forgotten, excluded, or neglected possibilities.

Another reason to separate the two stages is that they don't always occur together. There are cases of awareness that don't involve learning and others (like Movies) that do. Insisting that the two be treated identically seems, to me, unmotivated.

A general model of belief revision

In the discussion of Movies above, I brushed over an important question. I claimed that we need to separate out the stages of awareness growth and learning, but I didn't address how this learning takes place. How should we model agents who learn about the new propositions that they become aware of?

We must start by recognising that learning after (or, as part of) awareness growth cannot take place by Bayesian conditioning. When an agent Bayesian learns E, they update their beliefs from P to Q, such that $Q(X) = P(X|E)$, while holding their conditional probabilities fixed $Q(\cdot\ |X) = P(\cdot\ |X)$ for all $X \in \Omega$. But in an awareness growth case, E is a new proposition. Which is to say $E \notin \Omega$ and no conditional probabilities $P(\cdot\ |E)$ are defined. This is part of the very setup of awareness growth. So Bayesians who venture into these waters must give up their familiar comforts.

So, if not Bayesian learning, then what? We need a much more general model of belief revision, that doesn't assume priors involving the learned proposition and ideally one which can handle different kinds of learning: recall that in the examples above, the agent's learning involved fixing a conditional probability linking an old and new proposition. I will adopt a simple and general picture of belief revision that accomplishes these goals. It is inspired by how Richard Jeffrey models cases of

[15] Again note the problematic nature of the Relativity example in which LC is a postulate of SR. But imagine for the moment that SR were merely evidentially relevant to LC.

uncertain learning: instead of modelling what was learned as a proposition, Jeffrey proposes that we *describe the effects* of the experience on the agent, by stipulating their credences over a partition after the experience. He then provides a rule for generating a fully-specified, unique and coherent posterior credence, now called Jeffrey conditioning (Jeffrey 1983, 165).

Following the formalism of Dietrich, List, and Bradley (2016), I will think of a belief revision rule as a function, mapping an initial belief state and an experience to a final belief state. Let $\mathcal{P}$ be a set of possible belief states, $\mathcal{I}$ be a set of possible inputs or experiences, so that a belief revision rule maps $\mathcal{P} \times \mathcal{I} \to \mathcal{P}$. Belief states will be probability functions, or sets of probability functions, as before. As we're interested in belief revision following an extension to a new algebra, we will typically begin with a highly imprecise state. "Inputs" are taken to be very general, including straightforward observations, noisy signals, expert reports of various kinds. We therefore specify them *extensionally*, as the set of belief states that are consistent with the experience. A simple example: if I look out the window and see that it is cloudy, this input constrains my belief state to include only those in which it is cloudy outside my window.

Belief revision rules can be characterised by two conditions: Responsiveness and Conservatism (Dietrich, List, and Bradley 2016). Loosely, Responsiveness ensures that the final belief state "respects" the input, and Conservatism ensures that the belief revision changes *only* what is "required" by the input. This is captured by a Conservation condition, that specifies which parts of the prior belief state must be conserved by the revision.

Rules which obey these two conditions follow a pattern called perturbation-propagation.[16] First, as the rule is *Responsive* to the experience, the input will directly bring about a change in belief state: the perturbation. Second, the remainder of the belief state is adjusted to reflect the impact of the input; this propagation makes use of the perturbation and the parts of the initial state that are preserved by the *Conservatism* of the revision. Table 1 shows two common examples: Bayesian updating and Jeffrey updating. In the Bayes comes to have probability 1 in a proposition E. In the Jeffrey case, the agent comes to have new probabilities π_A for each element A of a partition $\mathbb{A}$. Note that the propagation step covers what is typically thought of as a "belief revision rule," such as updating by Bayesian conditioning.

[16] I take this term from Bradley (2017).

Table 1. Two belief revision rules

Rule	Perturbation	Propagation
Bayes	$Q(E) = 1$	$Q(X) = P(X\|E)$
Jeffrey	$Q(A) = \pi_A, \forall A \in \mathbb{A}$	$Q(X) = \sum_{A \in \mathbb{A}} P\,(X\|A)Q(A)$

In generalised belief revision theory, a *kind of experience* is matched with a *particular revision rule*. Dietrich, List, and Bradley (2016) characterise the class of Bayesian inputs as those experiences which constrain the agent's belief state to include only probability functions in which the probability of a specific proposition—the one the agent learns during the experience—is 1. We can similarly define the class of Jeffrey inputs, corresponding to Jeffrey updates. In each case, we can model this with a domain $\mathcal{D}$ that is a subset of the space of possible experiences and initial states: $\mathcal{D} \subseteq \mathcal{P} \times \mathcal{I}$.

Responsiveness consists in ensuring that the final belief state is in the set I, i.e., that it is compatible with the experience. Conservatism is harder to spell out. Each domain that Dietrich, List, and Bradley (2016) consider comes with a specification of what those experiences are "silent" on. This notion of silence is used to fill out the norm of Conservatism: put roughly, a belief revision rule should leave unchanged whatever the experience is silent on. Dietrich, List, and Bradley (2016) then prove characterisation results showing that for Bayes, Jeffrey and Adams updating there is a unique rule respecting Responsiveness and Conservatism and that in each case it is the rule referred to parenthetically above.[17]

In awareness growth cases, we don't want to restrict the kinds of information the agent may receive about the new possibilities, for this would restrict the applicability of the model. Agents may come to know that one event is more likely than another, or that two variables are independent, or they might learn the expected value for some variable. (van Fraassen (1981) says that learning the expected value for a variable is the most general kind of constraint on your probability function, and that others can all be framed as special instances of it.) So I want my theory of belief revision to be able to handle all of these report types.

The problem is that for these more general input domains, no unique belief revision rule is known. Put extensionally, the problem is that once we identify the

[17] Adams updating is a form of learning in which an agent comes to have new conditional probabilities over a partition. It was introduced by Richard Bradley (2005, 2017).

set of belief states that respect the input, we lack general rules for further refining this set. This may be easier to see by switching to an intensional definition: let us denote the information learned in the experience with a formula, ϕ_Q. Responsiveness tells us that the posterior credence function must respect this constraint: we want a Q that respects ϕ_Q. But a great many functions will typically do this! What we want is one which also fits the extended prior, P^+, in the right way. What is that way? Conservatism is meant to provide the answer: in the way that preserves as much of P^+ as possible while respecting ϕ_Q.

Specifying the Conservatism norm for a form of experience is a complex matter. The canonical examples mentioned above have a particularly nice form: each comes with a constraint and a *rigidity* condition which realises the Conservation condition. These conditions, summarised in Table 2, are necessary and sufficient conditions for updating according to the associated belief revision rule (Bradley 2017, 188–200).

Table 2. Conservatism conditions for Bayes and Jeffrey updates. $\mathbb{A}$ is a partition.

	Constraint	Rigidity condition(s)
Bayes	$Q(E) = 1$	$Q(\cdot\,|X) = P(\cdot\,|X), \forall X \in \Omega$
Jeffrey	$Q(A) = \pi_A, \forall A \in \mathbb{A}$	$Q(\cdot\,|A) = P(\cdot\,|A), \forall A \in \mathbb{A}$

For general learning (not in cases of awareness growth) the challenge is this: the agent has priors for all propositions in the domain; *some* of these are replaced by ϕ_Q and this process breaks the coherence of the function. In order to restore coherence, we need to ″fit″ the remaining parts of P with the bits of Q specified by ϕ_Q. But there are many ways of doing this. In a Bayesian learning case the agent's prior $P(E)$ is replaced by $Q(E) = 1$. Making only this change and keeping the rest of P would lead to incoherence. Coherence is restored by the propagation rule: set the rest of Q equal to P conditional on E.

But for more general kinds of inputs (e.g., learning that expectation of a random variable), we don't have Conservation conditions that produce such unique "kinematic" update rules. One option is to to do more work to identify what is conserved by each kind of experience (i.e., to identify Rigidity conditions like those in Table 2) in order to formulate a kinematic revision rule. Another option is to find the coherent function which obeys ϕ_Q and is closest to P by some measure of ″distance″. There is a large literature on using divergences between probabilities for epistemological or decision theoretic purposes: such methods are present in foundational work by de Finetti and Savage; van Fraassen (1981) and Diaconis and

Zabell (1982) utilise such methods; and they are present throughout the "accuracy" programme in epistemology including in work by Joyce (1998) and Pettigrew (2016).

This is a significant challenge for belief revision theory generally. But things are much simpler in awareness growth cases, where the learning we are interested in concerns the new awareness. That is because the constraint ϕ_Q concerns the *new propositions*. In the examples we have looked at, the agent comes to have certain posterior probabilities $Q(E_i)$ for a set of new propositions $\{E_i\}$, or conditional probabilities $Q(X_j|E_i)$ where the X_j propositions can be new propositions or (embeddings of) old propositions. But the agent had no priors involving the $\{E_i\}$ propositions, that's part of what it means to say that they underwent a growth of awareness. So there is no conflict with any part of P, or P's extension to the wider algebra P^+. Recall that P^+ is typically very imprecise, a set of many probability functions which match the demands of Rigid Extension on the embedding of the old algebra, but which differ on the (rest of the) new algebra. So we can simply further restrict this set by imposing the belief constraint ϕ_Q gained in the learning experience.

Here is our final belief revision recipe:

> **Generalised belief revision.** An agent has undergone rational awareness growth, taking them from initial awareness state Ω and belief state P, to new awareness state Ξ and rigidly extended belief state P^+, via an embedding h. They also undergo a learning experience characterised by ϕ_Q. The rational belief revision is to adopt posterior belief state Q: the set of credence functions on Ξ which coincide with P^+ on $h(\Omega)$ and which obey ϕ_Q elsewhere. (Again, this set will typically not be a singleton.)

Recall the Movies case. Here the "learning" experience had two effects: Q(Thriller | Low)=1, and Q(Low) > Q(High), representing your associations between Thrillers and basic language on the one hand, and your assumptions about the tastes of the cinema owner for simplistic language on the other. So long as we read these as names of propositions in the new algebra, these two constraints constitute ϕ_Q.

Nothing more is required. We form a new algebra, in which an independent language-level partition {Low, High} is introduced. There is a natural embedding, mapping Thriller to Thriller, and so on. The agent's prior is rigidly extended to the new algebra, which yields the preservation of ratios expected by RB. Then the extended beliefs P^+ are updated in line with ϕ_Q. The endpoint is the intuitive result—that shown in Steele and Stefánsson's diagram, Figure 6. But we get there in stages: first by determining the new algebra, then by applying Rigid Extension (new), and finally by carrying out the generalised belief revision procedure.

References

Alchourrón, C. E., P. Gärdenfors, and David Makinson. 1985. "On the Logic of Theory Change: Partial Meet Contraction and Revision Functions." *Journal of Symbolic Logic* 50: 510–30.

Bradley, Richard. 2005. "Radical Probabilism and Bayesian Conditioning." *Philosophy of Science* 72 (2): 342–64. https://doi.org/10.1086/432427.

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

Davey, B. A., and H. A. Priestley. 2002. *Introduction to Lattices and Order*. 2nd ed. Cambridge: Cambridge University Press.

Diaconis, Persi, and Sandy L. Zabell. 1982. "Updating Subjective Probability." *Journal of the American Statistical Association* 77 (380): 822–30.

Dietrich, Franz, Christian List, and Richard Bradley. 2016. "Belief Revision Generalized: A Joint Characterization of Bayes' and Jeffrey's Rules." *Journal of Economic Theory* 162: 352–71. https://doi.org/10.1016/j.jet.2015.11.006.

Earman, John. 1992. Bayes or Bust: A Critical Examination of Bayesian Confirmation Theory. Cambridge, MA: MIT Press.

Eels, Ellery. 1985. "Problems of Old Evidence." *Pacific Philosophical Quarterly* 66: 283–302.

Fraassen, Bas van. 1981. "A Problem for Relative Information Minimizers in Probability Kinematics." *The British Journal for the Philosophy of Science* 32 (4): 375–79. https://doi.org/10.1093/bjps/32.4.375.

Glymour, Clark N. 1980. *Theory and Evidence*. Princeton: Princeton University Press.

Jeffrey, Richard. 1983. *The Logic of Decision*. 2nd ed. University of Chicago Press.

Joyce, James M. 1998. "A Nonpragmatic Vindication of Probabilism." *Philosophy of Science* 65: 575–603.

Karni, Edi, and Marie-Louise Vierø. 2013. "'Reverse Bayesianism': A Choice-Based Theory of Growing Awareness." *American Economic Review* 103 (7): 2790–810. https://doi.org/10.1257/aer.103.7.2790.

———. 2015. "Probabilistic Sophistication and Reverse Bayesianism." *Journal of Risk and Uncertainty* 50: 189–208.

———. 2017. “Awareness of Unawareness: A Theory of Decision Making in the Face of Ignorance.” *Journal of Economic Theory* 168: 301–28. https://doi.org/10.1016/j.jet.2016.12.011.

Levi, Isaac. 1977. “Subjunctives, Dispositions and Chances.” *Synthese* 34: 423–55.

———. 1980. The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance. MIT Press.

———. 1991. *The Fixation of Belief and Its Undoing*. Cambridge, MA: Cambridge University Press.

Mahtani, Anna. 2020. “Awareness Growth and Dispositional Attitudes.” *Synthese*. https://doi.org/10.1007/s11229-020-02611-5.

Pettigrew, Richard. 2016. *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.

Schipper, B. C. 2015. “Awareness.” In *Handbook of Epistemic Logic*, edited by H. van Ditmarsch, J. H. Halpern, W. van der Hoek, and B. Kooi, 77–146. College Publications.

Shimony, A. 1970. “Scientific Inference.” In *The Nature and Function of Scientific Theories*, edited by R Colodny, 79–172. University of Pittsburgh Press.

Steele, Katie, and H. Orri Stefánsson. forthcoming. *Beyond Uncertainty*. Cambridge University Press.

Vallinder, Aron. 2018. “Bayesian Variations: Essays on the Structure, Object, and Dynamics of Credence.” PhD Thesis, London: London School of Economics and Political Science.

Wenmackers, S, and Jan-Willem Romeijn. 2016. “New Theory About Old Evidence.” *Synthese* 193 (4): 1225–50.

Katie Steele[1]

Why Time Discounting Should Be Exponential: A Reply to Callender[2]

According to Craig Callender (2020), the "received view" across the social sciences is that, when it comes to time and preference, only *exponential* time discounting is rational. Callender argues that this view is false, even pernicious. Here I endorse what I take to be Callender's main argument, but only insofar as the received view is understood in a particular way. I go on to propose a different way of understanding the received view that makes it true. In short: When time discounting is suitably conceived, the exponential form of the discounting function is indeed uniquely rational.

[1] School of Philosophy, Australian National University, Canberra, ACT 2600, Australia. *Email:* katie.steele@anu.edu.au

[2] Financial support by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences) is gratefully acknowledged.

1. Time's significance

The temporal position of goods matters to their value in a variety of ways. Consider the following stories about Margot and Amisha, which illustrate two ways in which the timing of eating oysters can matter.

> *Summertime Margot*. Margot loves oysters in the summertime when the air is warm and smelling of salt, and when she is on holidays, in a relaxed frame of mind. It is now the last week of her summer holidays. Oysters are worth a lot to her this week. But she will not sign up for weekly deliveries of oysters, as oysters next week, when she returns to her normal work routine, are hardly worth anything to her. Ditto for oysters the week after that.
>
> *Impatient Amisha*. Amisha also loves oysters. She loves them all year round. But she is impatient. It is now the last week of her summer holidays. Oysters are worth a lot to her this week. But she will not sign up for weekly deliveries of oysters, as oysters next week, given the wait, are not worth so much to her. Oysters the week after that, given the even longer wait, are worth proportionately less again.

There are many other stories that could be told in which the timing of a good matters to its value.

According to the "received view" of time discounting, however, some ways preferences relate to time are rational while others are not. Followed to the letter, the received view regards Amisha's preferences for oysters as rational but maintains that Margot's are irrational. Never mind that Amisha is merely impatient, whereas Margot has more substantial reasons for valuing oysters differently depending on their temporal position. What matters, roughly speaking, is that the effect of time distance on Margot's preferences for oysters is not steady or linear. For Margot, this week is qualitatively different to all other weeks. Amisha's preferences, on the other hand, correlate with how far in the future her eating oysters occurs. As a result, Amisha can be represented as having a *time discounting function* for oysters that has the approved *exponential* form, whereas Margot cannot be so represented; her time discounting function is non-exponential. This will be explained more fully in section 2 below.

Craig Callender presents a convincing case that it is not just odd but wrong to regard Margot and Amisha as having different status as far as rationality goes. Callender employs both *modus tollens* and *modus ponens* reasoning. That is, he works backwards—"The conclusion of the received view is wrong and so therefore its premises must be faulty"—and he also works forwards—"The premises of the received view are faulty and so the conclusion is not secured". I will outline the high points of

Callender's arguments in section 3, with reference to the stories of Margot and Amisha.

Like Callender, I regard both Margot and Amisha as rational, and so consider any account of time discounting that suggests otherwise to be mistaken. Moreover, I agree with Callender that one fairly salient way of reconciling the received view with Margot's rationality is unconvincing (section 4.1). I go on to propose, however, a subtly different refinement of the received view that Callender overlooks (section 4.2). It involves clarifying the conditions under which an agent's preferences can be represented in terms of a time discounted utility model. When these conditions are met, the agent is rational if and only if her time discounting function is exponential. Moreover, by way of conclusion (section 5), I suggest that it is really this refined version of the received view that one well known advocate of exponenttial discounting—the economist Robert Strotz—intended all along.

2. Inferring the discount function

On the "received view", the way an agent *discounts goods for time* can be inferred from her preferences over outcomes that differ in the temporal position of goods. More on how this works shortly. Let us first back up and put this talk of time, goods and preferences in the context of the standard account of rational choice: expected utility (EU) theory. We will see that EU theory itself leaves open how the timing of goods affects an agent's preferences. The received view about time discounting amounts to a supplement to EU theory.

On EU theory, an agent is rational just in case she can be represented as having a utility function, u, that satisfies the expected utility principle. According to this principle, her utility for risky options is the probability-weighted average of the utility of the possible realisations (or outcomes) of the option. This can be stated as follows (where A is a risky option, P is the agent's subjective probability function and o_i are the possible realisations or outcomes of A):

Expected Utility Principle. $u(A) = \sum_i P(o_i) \times u(o_i)$

The outcomes o_i are supposed to be ways that the world could be; at the limit of precision they are fully detailed world histories. But little is said about the agent's preferences over these world histories except that they are well ordered. Why an agent prefers one outcome or world history to another is not analysed; it is simply a black box.

There have been various suggested supplements to EU theory, however, by way of adding further structure to outcomes and associated rationales for comparing them.

The general move is to depict outcomes as having specified properties or attributes that each contribute some fixed value to the outcome. For instance, so-called *multi-criteria decision analysis* is a family of approaches to distinguishing and aggregating different sources of value in an outcome. Dietrich and List (2013) have finessed this idea in proposing a single, abstract model that accommodates all the ways in which the value of an outcome may be decomposed into the value of its constituent properties.

The economists' depiction of outcomes as temporal streams of consumed goods can be seen as a specific version of these general approaches to structuring outcomes. On this depiction, outcomes are comprised of temporally-indexed bundles of goods; they are *consumption streams*. Koopmans (1960) for instance, represents outcomes as $x = (x_1, x_2, x_3, \dots)$, where the x_t are vectors of goods. That is, the vector $x_t = (x_{t1}, x_{t2}, \dots x_{tn})$ lists the non-negative amounts of each of the n goods to be consumed at time t. Depending on the structure of the agent's preferences over these consumption streams, her preferences can be represented by a utility model of greater or lesser simplicity.

Early on, pioneering economists like Ramsey (1928) and Samuelson (1937) singled out, for its nice features, the *exponential time discounted utility model* for evaluating consumption streams. On this model, an agent at time τ has utility, u_τ, for a consumption stream, x, as follows (where t is the time variable, $D_\tau(t-\tau)$ is the agent's time discounting function over time distances given by $t-\tau$, and u is her utility function over vectors of goods, x_t):[3]

Time Discounted Utility. $u_\tau(x) = \sum_{t=\tau}^{t=\infty} D_\tau(t-\tau) \times u(x_t)$

Moreover, the agent's time discounting function, $D_\tau(t-\tau)$, has the following form (where ρ is the so-called discount rate):[4]

Exponential Time Discounting. $D_\tau(t-\tau) = (\frac{1}{1+\rho})^{t-\tau}$

That is, each unit of time distance from the present, τ, reduces the value of a bundle of goods by a fixed proportion. Koopmans (1960) showed the full set of conditions under which an agent's preferences over consumption streams may be represented in terms of an exponential discounted utility model. But, as Callender points out, the main property that distinguishes exponential from other forms of time discounted utility is Stationarity (on which more below).

[3] The expression here allows for an infinite consumption stream that begins at the present time, τ, but it can be adjusted to incorporate past times or to have an upper limit on the end time of the consumption stream.

[4] The expression is for discrete time. The continuous counterpart is $e^{-\rho(t-\tau)}$.

We need not consider full consumption streams in order to understand the time discounting of Margot and Amisha. A simpler model will suffice in which outcomes can be represented as pairs (x, t), denoting that some bundle of goods x is consumed at time t.[5] In fact, we can go yet simpler, and assume that the agent only cares about the consumption of one type of good. (In the case of Margot and Amisha, it is oysters.) In this case, outcomes can be represented as (s, t), where s stands for the quantity of the good that is consumed and t is the time at which it is consumed.[6]

The property of Stationarity mentioned above can be spelled out for preferences over the simple outcomes (s, t).[7] (It is spelled out in terms of the indifference relation from the perspective of time τ, denoted $\approx_\tau$.) Consider outcome quantities s, r whose values are real numbers, and times t, t', and time distances $\Delta_1, \Delta_2 \geq 0$. The set of preferences satisfies Stationarity if:

Stationarity. $(s, t + \Delta_1) \approx_\tau (r, t + \Delta_2) \Leftrightarrow (s, t' + \Delta_1) \approx_\tau (r, t' + \Delta_2)$

We see that this condition requires that preferences from the perspective of time τ over temporally-indexed goods depends only on the nature of the goods themselves and the time distance between them, not *where* they are in time on a calendar, so to speak.

It is because Amisha's preferences over the outcomes (s, t) satisfy Stationarity while Margot's do not that Amisha can be represented as having an exponential time discounting function with respect to these outcomes, whereas Margot cannot be so represented.

First consider Margot. Let's say we know (filling in the story a little) that, now at time τ, $(n, 0) \succ_\tau (n, 1)$ and $(n, 1) \approx_\tau (n, 2)$ where n is any given number of oysters, and time is measured in terms of weeks, such that $t = 0$ stands for the period up until one week has passed. Margot does not satisfy Stationarity, since she prefers n oysters now to the same number of oysters in one week's time, but this pattern of preference does not hold when the timeline shifts; it does not hold for n oysters next week compared to the same number one week after that. Any time discounted model of Margot's preferences over the outcomes (n, t) will thus not be one in which the time discounting function is exponential. For instance, a candidate model is one

[5] Lancaster (1963) appeals to outcomes of this form in his investigation of time discounting. As he points out, one can think of these pairs as special cases of consumption streams, where no goods are consumed (or else the status quo is maintained) for all times except the stated time t when bundle x is consumed.

[6] For various reasons, Fishburn and Rubinstein (1982) and Halevy (2015) use this very simple kind of outcome in their respective investigations of time discounting.

[7] Like Callender, I appeal to Halevy's (2015) definition of Stationarity (although with some differences in notation); In the next section I will appeal too to Halevy's (2015) definitions of the further conditions of Consistency and Invariance.

where her time discount factor for week 0, $D_\tau(0)$, is one, while her time discount factor for the two subsequent weeks is $w < 1$. That is, $D_\tau(1) = D_\tau(2) = w$. If u_M is Margot's utility function over varying amounts of oysters, her utility at time τ for the outcome $(n, 1)$ is $1 \times u_M(n)$ whereas her utility at time τ for both $(n, 1)$ and $(n, 2)$ is $w \times u_M(n)$.

Now consider Amisha. We know that $(n, 0) \succ (n, 1) \succ (n, 2)$. Specifically, Amisha prefers n oysters during the first week to n oysters during the second to the same extent as she prefers n oysters during the second week to n oysters during the third week. Her preferences satisfy Stationarity: what determines her preference for n oysters is the distance between the temporal locations of these oysters and not where the locations lie on the calendar. So if we were to model Amisha's preferences at the current time τ in terms of a time discounted utility model, her discounting function would be exponential in form. For instance, a candidate model is one where $D_\tau(0) = 1, D_\tau(1) = 0.5$ and $D_\tau(2) = 0.25$. In this way $\frac{D_\tau(2)}{D_\tau(1)} = \frac{D_\tau(1)}{D_\tau(0)}$. If u_A is Amisha's utility function over varying amounts of oysters, her utility at time τ for outcome $(n, 0)$ is $1 \times u_A(n)$, her utility at time τ for outcome $(n, 1)$ is $0.5 \times u_A(n)$ and her utility at time τ for outcome $(n, 2)$ is $0.25 \times u_A(n)$.

3. Callender against the received view

On the received view, only exponential time discounting is rational. But what exactly does that mean? Callender suggests it means that agents like Margot (as compared to agents like Amisha) are irrational. In that case, the received view must be something like the following:

Received View. *If an agent's preferences at some time τ can be represented in terms of a time discounted utility model, she is rational if and only if the form of her time discounting function is exponential.*

Whether or not this is really the dominant way of thinking about time discounting is, of course, a sociological matter (hence Callender quotes various economists, psychologists and evolutionary theorists in support of his attribution). Here I will simply take for granted that Callender has accurately represented the views of social scientists on discounting.

Callender goes on to argue against the received view, understood as above (referred to from on now as Received View). For starters, he finds the conclusion that agents like Margot are irrational to be absurd. Moreover, by drawing on an insightful theorem by Halevy (2015), he shows that the conclusion rests on shaky foundations. This section retells, using the stories of Margot and Amisha, Callender's important criticisms of Received View. This sets the stage for the next section, which explores how Received View may yet be slightly adjusted so that it is defensible.

3.1 Bad conclusion

Let us begin with the absurdity of claiming that Margot (unlike Amisha) is irrational. First, to call someone irrational suggests that they are confused—that their preferences and motivations are in some sense incomprehensible. But that is not true of Margot. Her preferences, since they do not involve mere impatience, are arguably even more comprehensible than Amisha's. More generally, there are a variety of reasons why the temporal position of goods matters. After all, time is correlated with many things—ageing, the seasons, an evolving climate, technological innovation, all sorts of cultural trends—that serve as the background to goods and affect how and the extent to which these goods are enjoyed. It would be very surprising if the impact of these various temporal factors (as we might call them) could in all cases be expressed as exponential discounting of the goods in question. (And more surprising again if the exponential rate of discounting was the same for all goods, as some models suggest.) That is, there is nothing about exponential discounting that makes it the more obvious and explanatory way in which the temporal position of a good affects an agent's preferences.

Moreover, it is not as if exponential rather than non-exponential discounting leads to personal ruin. Callender notes that any suggestion to the contrary is a prime example of normative views clouding one's empirical observations, something that should, as far as possible, be avoided. (He thinks this theory-ladenness has not been avoided, nor well appreciated, in the literature. That is why he refers disparagingly to the received view as "a package": in addition to the normative claim about exponential discounting, there are psychological and evolutionary findings that purportedly support the normative claim, showing the problems that befall non-exponential discounters and offering complicated explanations for why many people might have this defect.) As Callender points out, both exponential and non-exponential discounters may have discounting curves that lead to later regret, not to mention financial and social poverty. After all, exponential discounting curves can be as steep as you like, effectively recommending that one "live like there is no tomorrow", placing very little value on goods to be enjoyed at future times. So if due concern for one's future self is the issue, exponential discounting is not the answer. Rather, the answer is no discounting at all, at least for the goods that are crucial to well being at a time.

3.2 Bad premises

So what exactly is the problem with some, but not all, forms of time discounting? As Callender points out, it was Strotz (1956) who made explicit what economists regard to be the real sin: *dynamic inconsistency*. The idea is that one's preferences over goods consumed at a given time should not depend on one's current vantage point. For otherwise, one would be vulnerable to exploitation. For instance, let's say that I currently do not place a high value on having oysters two weeks hence. So I am willing to sell the rights to oysters at that time for a small price. Suppose too that once two weeks have passed, I like oysters a great deal and am willing to pay a lot for them. Someone could make a profit by initially buying my two-weeks-in-the-future oysters at a low price and then later selling them to me at a high price. In general, whenever my preferences change, I can be exploited in this way. If I can predict with certainty that my preferences will change, the exploitative trades or bets can be engineered (a so-called "Dutch book") so that I already know in advance that I will suffer a loss. That seems bad, and a problem for any form of time discounting that permits it.

The property in question (or rather its converse) can be referred to as Consistency, and formally defined as follows (where outcomes are represented as before, t and t' are times where goods are located, $\approx_\tau$ is indifference from the perspective of time τ and $\approx_{\tau'}$ is indifference from the perspective of time τ'):

Consistency. $(s,t) \approx_\tau (r,t') \Leftrightarrow (s,t) \approx_{\tau'} (r,t')$

We see that this condition specifies that one's preferences for outcomes do not depend on one's vantage point, or the time at which one has the preferences.

Let us grant that rationality in the dynamic setting requires Consistency.[8] What Callender draws attention to is that, assuming already a time discounted utility model, the Consistency condition does not entail Stationarity (and with it, exponential time discounting), as many take Strotz's (1956) theorem to show. Nor does Stationarity entail Consistency. This is evident just by looking at the form of the two conditions. Stationarity concerns preferences at a particular time (cashed out in terms of the indifference relation $\approx_\tau$ that is indexed to the time τ). Consistency, on

[8] Callender does not in fact grant this. He cites well-known criticisms of (diachronic, in particular) Dutch book arguments. Some of these criticisms can be disarmed by emphasising that what matters (and what is the target of Consistency) are one's *plans* for whether and how one will change one's preferences over time. These plans are faulty or irrational when one can *anticipate* sure loss. As such, it does not matter whether or not one spontaneously changes one's preferences over time in an unanticipated way. No-one can guarantee a Dutch book against you if you change your preferences in unforeseen ways. So the conviction that one should be able to spontaneously change one's preferences at any given time is not reason to reject diachronic Dutch book arguments, nor, ultimately, Consistency.

the other hand, concerns the relationship between preferences at different times (how $\approx_\tau$ relates to $\approx_{\tau'}$). Something else is needed to bridge the logical gap between Stationarity and Consistency.

The something else is an assumption in Strotz's theorem that seems to have gone unnoticed.[9] Halevy (2015) draws attention to the suppressed assumption, dubbing it *Invariance*. It can be formally stated as follows (where the notation is as per the statement of the previous conditions):

Invariance. $(s, \tau + \Delta_1) \approx_\tau (r, \tau + \Delta_2) \Leftrightarrow (s, \tau' + \Delta_1) \approx_{\tau'} (r, \tau' + \Delta_2)$

Invariance requires that one's preferences with respect to time have the same form, whatever position on the calendar one currently occupies. That is, what matters is how far away in time the goods are from one's current vantage point.[10] Halevy (2015) goes on to make clear the logical role that Invariance plays. He reveals, in Callender's words (p. 13), "a beautifully simple relationship amongst the three temporal conditions, Consistency, Stationarity, and Invariance, namely: Any two implies the third."

The significance of bringing Invariance to light is that we see that one does not need to be an exponential discounter (satisfying Stationarity) in order to be immune from Dutch books (satisfying Consistency). One just needs to violate Invariance in a way that preserves Consistency. Moreover, this seems a perfectly reasonable approach to time discounting. Indeed, Callender notes that empirical studies suggest that many and perhaps most people who violate Stationarity in fact violate Invariance rather than Consistency.[11]

Margot is plausibly one such person. As the story goes, the reason Margot loves oysters this week is that she is still on summer holidays and so is in the mood for them. Given her prediction that she will not be in the mood for oysters next week or the week after, when she is back at work, she does not currently value consuming oysters in those weeks. That's as far as our story of *Summertime Margot* goes—it concerns only Margot's preferences at a single time. But the most plausible extension of the story is one where Margot sacrifices Invariance for Consistency.

[9] Note that this can hardly be regarded the fault of Strotz, who does not try to hide any of the assumptions supporting his (1956) theorem. I return to this point in the concluding section.

[10] This is subtly different to the Stationarity condition, which required that preferences over outcomes depend only on the nature of the outcomes and the time distances between *the outcomes*. Invariance on the other hand says that the preferences over outcomes depend only on the nature of the outcomes and their time distances from the *agent*.

[11] Callender cites the empirical results of Halevy (2015) and Janssens et al. (2017).

Consider:

> *Summers-end Margot*. A week has now passed and Margot is back at work. Just as she predicted, she does not particularly value oysters at this time. Nor does she particularly value having oysters next week, since she predicts she will then be similarly blasé about oysters. Ditto for the foreseeable weeks after that while her work routine continues.

(Note that this story and the discussion below are framed in terms of Margot's actual attitudes once she returns to work. But it could be retold in terms of Margot's predictions or *plans* for what her attitudes will be at this later time.[12]) The combination of preferences in *Summertime Margot* and *Summers-end Margot* seem entirely reasonable. They satisfy Consistency, since Margot does not change her preferences for the dated oysters over time. From the start, she did not much value having oysters in either week 1 or week 2, and this attitude remains unchanged as time passes and week 1 becomes the present. Margot thus cannot be Dutch booked. She maintains Consistency, however, by violating Invariance. At the earlier time, she loves oysters *now,* but not so much *later*. The effects of time distance are not invariant, however, with the passing of time. When she returns to work, it is not the case that Margot loves oysters *now,* but not so much *later*. Rather, she neither loves oysters *now* nor *later*.

We see that Margot satisfies Consistency and thus cannot be Dutch booked, but is there nonetheless something bad about her pattern of preferences? Callender argues that Received View can only be saved if Invariance is given some normative defence. He suggests that, at best, Invariance might be defended on the grounds of non-arbitrariness. Roughly, it would be arbitrary and thus irrational for one's time preferences to depend on one's current position in calendar time. But he goes on to say that this line of defence is wanting. Indeed, it is hard to see how such a line could be made convincing in the face of stories like Margot's.

It is worth noting too that "non-arbitrariness" is a rather ambiguous notion in this context. Callender quotes such luminaries as Adam Smith, Henry Sidgwick and John Rawls as advocates of non-arbitrariness of time preferences. But it is a stretch to claim that these scholars argue specifically for non-arbitrariness *in the sense of Invariance*. After all, each of Stationarity, Consistency and Invariance encode a non-arbitrariness *of sorts*. In fact, the quotes Callender appeals to all suggest a stronger form of non-arbitrariness than any of these three conditions. They can be interpreted as prescribing that the temporal position of a good should make no difference

[12] Recall footnote 8.

to an agent's preferences at all. This amounts to no time discounting of goods whatsoever, at any time. Many would think that this is far too restrictive and thus not credible. In any case, it is not a carefully targeted argument for Invariance.

Callender concludes that there is no normative motivation for Invariance, and so there are insufficient grounds to accept Received View. He thus claims that rationality does not require that time discounting be exponential. In fact, he suggests that once the pillar of Invariance is removed, the entire time discounted utility model may come crashing down. Perhaps, after all, it does not make sense to model an agent's outcomes and preferences in a way that makes time an independent, separable factor, especially once we move beyond simple settings involving a single good, as per the stories of Margot and Amisha. Rather, the timing of a good may be better seen as integral to its description, since it may impact on the enjoyment of the good in a way that is idiosyncratic to the good in question.

4. The received view redux

While Callender is right to insist that agents like Margot are not irrational, he is too quick to dismiss the notion that exponential time discounting has a special claim to rationality, not to mention the usefulness of the time discounted utility model. This section explores how the received view may be refined; I argue that it should be adjusted to reflect stricter conditions for when an agent's preferences may be modelled in terms of a time discounted utility model. In the conclusion I will suggest that these were the conditions Strotz originally put forward.

4.1 First pass: focus on *genuine* time discounting

Before getting to the main event, let us consider one salient way of making the conditions stricter for employing a time discounted utility model; a strategy that Callender rightly regards as flawed. This is to insist that a time discounted utility model of an agent's preferences is apt only if it separates out the role truly played by time itself, as opposed to some other properties correlated with time. The idea is that rationality requires only that *genuine* time discounting, in the sense just described, need have exponential form. The view can be spelled out as follows:

Received View+. *When an agent's preferences at some time τ are represented by a genuine time discounted utility model, she is rational if and only if the form of her time discounting function is exponential.*

The advantage of Received View+ is that it allows one to say that apparent counterexamples to the irrationality of non-exponential time discounting, like the case of Margot, simply are not genuine cases of time discounting. After all, it is not time itself that makes a difference to Margot's evaluation of oysters. Rather, it is the presence or absence of warm salty air and Margot's levels of stress that make for better and worse oyster eating. So an appropriate model of Margot would require a more complex description of the outcomes. There would need to be at least two types of goods: say, "holiday oysters" and "routine oysters".[13] Once the goods that matter to Margot are fully accounted for in these terms, we see that there is no further effect, owing purely to temporal position, on Margot's preferences. She does not exhibit any discounting for time whatsoever. (And so she satisfies, in a trivial way, all three conditions of Invariance, Consistency and Stationarity.)

While this may seem a natural enough move in the case of Margot, the idea that we would be able to distinguish what is and what is not purely an effect of time on preferences seems implausible. Even the phenomenon of impatience, which seems to be the paradigmatic effect of time itself on preferences, could be construed in non-temporal terms. What really matters to the agent is not temporal position, but rather the feelings of frustration due to waiting that accompany the good, one might say. In general, it would be very difficult, perhaps impossible, to locate the line between the effects of time on an agent's preferences versus the effects of other properties correlated with time. And even if this line could be located, this would trivialise exponential discounting as a norm of rationality. In many cases, as per Margot, the agent's time discounting would have exponential form, but only trivially, in that she does not discount for time at all. Moreover, she may yet violate Consistency for other reasons.

4.2 The way forward: focus on *whatever counts as* time discounting

Rather than try to pin down the real or *genuine* effects of time on preference, we can focus simply on what is a *useful* characterisation of the effects of time on preference. By way of spelling this out, I draw attention to the different roles that may be played by the Invariance condition. Callender takes Invariance to be an ostensive norm of rationality, albeit not a very convincing one. One way to maintain that it is a convincing norm is to argue, as per above, that apparent counterexamples do not truly undermine the norm—they trade on a faulty depiction of outcomes. But there is another reading altogether of the Invariance condition. As opposed to a norm, it may

[13] Margot's outcomes would then be represented $(< s, r >, t)$, where s is the quantity of holiday oysters and r is the quantity of routine oysters.

serve rather as a fixed point or assumption that allows one to see how distance in time affects an agent's preferences. The idea is that once we model an agent as having preferences over outcomes that are Invariant, then we can read off from the model the role that time itself plays in her preferences. After all, the effects of temporal distance, if they really are simply effects of temporal distance, do not depend on the agent's present calendar location. The agent discounts for temporal distance in whatever fashion her Invariant preferences over time suggest.

The model of Margot that we obtain when we *assume* Invariance is the same one described above. But the way we arrive at the model is different. We do not ask what are the properties of outcomes that must be accounted for in describing Margot's preferences that can be distinguished, objectively speaking, from time itself. Rather, we search for a way of describing outcomes so that Margot satisfies Invariance. (There will be such a description of Margot's preferences just in case she is, or plans to be, a stable person over time. Unstable agents who lack any systematicity in their preferences at different times are arguably not very interesting topics of study.) In Margot's case, the obvious way to do this is to discriminate holiday oysters and routine oysters, as before. Then Margot satisfies Invariance in a trivial fashion: the temporal distance of consuming oysters, at any given time, plays no role in her preferences. Her time discounting is thus rational because she satisfies, trivially, Consistency, and therefore Stationarity too.

There are other stories that better demonstrate how this approach—assuming Invariance—is distinct. Consider:

> *Stockpile Sarita*. Sarita loves oysters all year round, but she likes them better the fresher they are. She can store an evening's worth of oysters in her freezer. The freezer keeps the oysters pretty well, but their freshness nonetheless deteriorates linearly with time. As such, Sarita currently values consuming the oysters in proportion to how far away in time this would occur—when she would be removing them from the freezer and eating them. Moreover, she predicts that at any later date, she will similarly value consuming the oysters in proportion to how far away in time this would occur.

If we aimed to disentangle the genuine effect of time distance on Sarita's preferences, we would end up with a trivial model of her time discounting, as per Margot. That is because it is clearly not really temporal distance *per se* that matters to how Sarita values consuming the oysters taken from her freezer. Rather, it is their relative freshness, which is simply a physical attribute of the oysters. We would conclude that the apt model of Sarita's preferences is one that distinguishes all the different grades of oysters in terms of their freshness. There would be no further aspect of Sarita's preferences to be explained by distance in time itself.

One could model Sarita as caring nothing for temporal distance itself at all points in time; this is one way in which she can be modelled as satisfying Invariance. But that is to unnecessarily forgo the benefits of a simple and elegant model of Sarita's preferences. We can just as well model Sarita as having a non-trivial time discounting function. Sarita satisfies Invariance when her outcomes are described simply as a given amount of oysters at some temporal distance, represented (s, t). Moreover, the story indicates that at all times temporal distance matters to her in a *linear* fashion (satisyfing Stationarity). Sarita's preferences can thus be depicted in terms of a time discounted utility model where the time discounting function is exponential. So we see, in a very simple and vivid way, that she is Consistent and therefore arguably rational.

This approach sidesteps unnecessary debates. By way of another example, consider a slight modification of the story *Impatient Amisha*. In this case, Amisha's impatience is such that she greatly values having oysters in the near future, while having them at more distant times has more or less the same low value to her. Let us assume that at future times too, her impatience takes the same form. It would be silly to agonise over what is Amisha's *genuine* relationship with time. Does her impatience mean that time really matters to her? Or is impatience, too, reducible to other properties of outcomes, e.g., feelings of frustration? Consequently, is the right model of Amisha one in which she exhibits non-exponential discounting, or rather one in which she does not discount for time at all? This question is a mere distraction. All that matters is whether Amisha satisfies Consistency. It is clear that she does not, since *there exists* a model of her preferences that satisfies Invariance and yet violates Stationarity (such that her time discounting is non-exponential). Therefore her preferences violate Consistency.

This brings us back to the received view about time discounting. My amended version of it can be stated as follows:

Received View*. *If an agent's preferences at all times τ within a relevant period of time can be represented in terms of a time discounted utility model, she is rational if and only if the form of her time discounting function is exponential.*

For an agent's preferences at different times to be modelled by *a* time discounted utility model—the requirement in Received View*—she must satisfy Invariance. Arguably, a time discounted utility model is indeed inherently temporally extended in this way. That is, it makes no sense to say that an agent's preferences at a single time can be represented by a time discounted utility model. Only when we consider what is the common form of an agent's preferences at multiple times can we see what role distance in time, rather than calendar time, plays. For a stable agent at least, distance in time has the same effect, whatever the agent's position in calendar time. And it is the effect of distance in time that is usefully captured by the discounting function in a time discounted utility model. Furthermore, on pain of irrationality, this discounting function must have exponential form.

5. Conclusion

By way of conclusion, I note that Strotz himself proposed something along the lines of Received View*. While he does not explicitly assume Invariance, it is implied by his insistence that a time discounting function capture the role that time distance (and not calendar time) plays in the preferences of an agent. Indeed, Strotz states in the introduction of his paper (1956, 165): "What is crucial to all this is that the discount applied to a future utility should depend on the time-distance from the present date and not upon the calendar date at which it occurs."

Moreover, Strotz takes pains to point out in a long passage (1956, 167–8) that his model can accommodate the effects of calendar time too (such that whatever the nature of an agent's preferences, so long as she has a certain stability, she can be modelled by an Invariant time discounted utility model). Instead of simply fine-graining the description of goods to account for calendar time (e.g, holiday oysters versus routine oysters), Strotz pursues a slightly more elegant model. He makes the agent's timeless utility function for the bundle of goods at some calendar time depend not just on the vector of goods but also on the calendar time. So where in the expression for Time Discounted Utility (refer back to section 2) there appears $u(x_t)$, Strotz has instead $u(x_t, t)$. This allows his model to incorporate calendar time as a separable aspect of a bundle of goods that nonetheless has an effect on these goods that is timeless; it does not depend on distance in time relative to the agent.

Strotz implies that it falls on the modeller to determine a time discounted utility model that is suitably representative of an agent as she moves through time. That is the spirit of Received View*. On this approach, exponential time discounting is, after all, privileged. It is the uniquely rational form of discounting in that exponential discounters alone satisfy Consistency.

References

Callender, C. 2018. 'The Normative Standard for Future Discounting', *The Australasian Philosophical Review,* forthcoming.

Dietrich, F. and C. List. 2013. 'A reason-based theory of rational choice', *Noûs,* 47(1): 104–134.

Fishburn, P. and A. Rubinstein. 1982. 'Time Preference', *International Economic Review,* 23: 677–694.

Halevy, Y. 2015. 'Time Consistency: Stationarity and Time Invariance', *Econometrica,* 83: 335–382.

Janssens, W., Kramer, B., and L. Swart. 2017. 'Be Patient When Measuring Hyperbolic Discounting: Stationarity, Time Consistency and Time Invariance in a Field Experiment', *Journal of Development Economics,* 126: 77–90.

Koopmans, T. C. 1960. 'Stationary Ordinal Utility and Impatience', *Econometrica,* 28(2): 287–309.

Lancaster, K. 1963. 'An Axiomatic Theory of Consumer Time Preference', *International Economic Review,* 4(2): 221–231.

Ramsey, F. P. 1928. 'A Mathematical Theory of Saving', *The Economic Journal,* 38(152): 543–559.

Samuelson, P. A. 1937. 'A Note on Measurement of Utility', *The Review of Economic Studies,* 4(2): 155–161.

Strotz, R. H. 1956. 'Myopia and Inconsistency in Dynamic Utility Maximization', *The Review of Economic Studies,* 23(3): 165–180.

Nicholas Lawson[1] & Dean Spears[2]

Population Externalities and Optimal Social Policy[3]

If fertility is not chosen in a socially optimal way, and if policies to directly target fertility are ineffective or politically infeasible, then public policies that affect fertility could have important welfare consequences through the fertility channel. We refer to these effects as population externalities, and in this paper we focus on one important variable that may have a causal impact on fertility: the education of potential parents. If increased education causes families to have fewer children, then a government would want to increase college tuition subsidies in the presence of environmental externalities such as climate change, to indirectly discourage families from having children who will generate future environmental costs. Alternatively, if fertility is inefficiently low, due to imperfect parental altruism for example, governments will want to lower tuition subsidies to encourage child-bearing. We present a simple model of the college enrollment decision and its fertility impacts, and show that such

[1] Department of Economics, University of Quebec in Montreal. lawson.nicholas@uqam.ca

[2] Economics Department and Population Research Center, University of Texas at Austin; Economics and Planning Unit, Indian Statistical Institute - Delhi Centre; IZA; Institute for Futures Studies, Stockholm; r.i.c.e. (www.riceinstitute.org). dean@riceinstitute.org

[3] Spears' research was supported by grant K01HD098313 and by P2CHD042849, Population Research Center, awarded to the Population Research Center at The University of Texas at Austin by the Eunice Kennedy Shriver National Institute of Child Health and Human Development. Financial support for the Climate Ethics project by Riksbankens Jubileumsfond is gratefully acknowledged. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would like to thank participants of the CIQSS Applied Micro Day 2019 in Montreal and the Conference on Population, Social Welfare, and Climate Change in Austin for their comments. Any errors or omissions are the responsibility of the authors.

population externalities are quantitatively important: the optimal subsidy increases by about $5000 per year with climate change, and decreases by over $7000 per year with imperfect parental altruism. Our paper demonstrates how public economics can incorporate population externalities, and that such externalities can have significant impacts on optimal policy.

1. Introduction

A wide variety of variables can affect both a family's decision to have children and their desired number of children. Clearly, there is considerable heterogeneity in the fertility preferences of families, but a number of economic variables can enter the decision, including macroeconomic conditions, the employment status of the potential parents (Currie and Schwandt 2014), public monetary fertility incentives such as baby bonuses (Milligan 2005), family policies such as parental leave and benefits (Hyatt and Milne 1991; Phipps 2000; Cannonier 2014), and even programs that are not targetted at fertility, such as social assistance (Robins and Fronstin 1996).

In particular, one important variable that may have a causal impact on fertility is the education level of the parents: a number of papers have found significant negative effects of parental education on fertility in developed countries, including (León 2004), (Amin and Behrman 2014), (Lavy and Zablotsky 2015), and (Fort, Schneeweis, and Winter-Ebmer 2016).[4] This is not a universal finding in the literature,[5] but the weight of evidence of recent empirical studies for the U.S. generally supports the conclusion that education has a negative causal effect on fertility.

However, the fact that a number of policies and policy-related variables could impact fertility doesn't necessary imply a welfare impact. If fertility is chosen by households in a socially optimal way, then there are no population externalities from such policies, due to a standard envelope condition: even if subsidizing education (for example) leads to a more educated population that chooses a lower level of fertility, if households choose that fertility level optimally given their level of education, the derivative of welfare with respect to fertility is zero and there are no first-order welfare consequences of the resulting change in fertility.

But what if fertility is not chosen in a socially optimal way? It is not hard to imagine mechanisms that could lead to an inefficiently high or low level of fertility – and thus to an inefficiently large or small population – and we focus on two in this paper: environmental externalities such as climate change, and imperfect altruism on the part of parents who do not fully internalize the future utility of their potential children.

[4] The latter paper finds a reduction in total fertility for the U.K. but not for continental Europe. Some other papers focus on timing of fertility and find that education causes a delay (Black, Devereux, and Salvanes 2008; Tropf and Mandemakers 2017; James and Vujić 2019), though perhaps no overall effect on total fertility (Monstad, Propper, and Salvanes 2008).

[5] Some papers find no effects at all, such as (Kan and Lee 2018) for Taiwan, and (Fort, Schneeweis, and Winter-Ebmer 2011) even finds a positive effect, but their study focusses on low-skill individuals affected by compulsory schooling laws for which positive income effects from increased education may outweigh negative substitution effects and lead to higher fertility.

Environmental externalities are relevant for questions of population because each person born adds to the amount of pollution produced, imposing social costs on the entire population. Several papers, including (Harford 1997), show that a standard pollution tax, or any other instrument that targets emissions per person, is ineffective in internalizing the population externality if the revenues from the policy are used in some way to benefit the population, include distributing them lump-sum to each individual. Indeed, (Harford 1997) shows that the optimal policy includes a "tax per child equal to the discounted present value of all pollution taxes that the child and her descendants would pay"; in the absence of such a child tax, fertility will tend to be inefficiently high, and the title of (Harford 1998) refers to this as "The Ultimate Externality." We focus in this paper on the specific and profoundly important case of climate change.

In the opposite direction, the field of population ethics suggests a possible reason why fertility could be inefficiently low: parents may not fully account for the future utility of their potential children when making the decision to have children or not. That is, parents may be imperfectly altruistic towards their children, and if the social planner wants to maximize the discounted stream of present and future utility from all generations – rather than just maximizing the parents' utility – the planner will place more weight on the utility of future children than their parents do, suggesting that fertility will be inefficiently low.[6]

If either (or both) of these arguments are correct, fertility and population levels may differ considerably from the socially efficient allocation. In such a case, the standard public economics response would be to attempt to correct the inefficiency at the source, such as with a subsidy or tax on childbearing, so as to ensure that parents face the efficient "price" of having a child. However, it is hard to alter fertility in rich countries; numerous studies summarized in (Gautier 2007) have found that while fertility subsidies may affect the timing of births, the overall effect on total completed fertility is considerably smaller, and (Blanchet and Ekert-Jaffé 1994) estimated that the total effect of fertility policies across various countries was about 0.2 children per woman. This implies that correcting any distortion to fertility using directly targetted policy could be quite difficult; in other words, if fertility responds slowly to subsidies and taxes, the optimal subsidy or tax could be infeasibly large.

In such a second-best world, we need to consider the impacts of other policies on fertility: if a policy causes the privately-optimal level of fertility to increase, for example, this would generate added welfare gains in the case of imperfect altruism, and added welfare losses in the case of environmental externalities. In both cases,

[6] An alternative mechanism leading to insufficient fertility could be market size effects on productivity and innovation: a larger population could generate more innovation in technologies that can benefit everyone, introducing a positive population externality.

various social policies could offer an indirect way to target fertility and population, and these welfare impacts should be considered when evaluating an optimal policy.

If education has a negative causal effect on fertility, as we will assume following the work of (Amin and Behrman 2014), then college tuition subsidies represent one such policy: if a higher tuition subsidy – or lower net costs of college – encourages more young individuals to become more educated, and if that in turn leads to reduced fertility, we may want a considerably larger subsidy than we otherwise would if we are concerned about climate change, whereas we may want a lower subsidy (and maybe even an education tax) if we are concerned about imperfect altruism.

We will consider this question in our paper, basing our analysis on (Lawson 2017), who considered optimal college tuition subsidies in the presence of fiscal externalities (positive effects of education on future tax revenues through higher incomes) and liquidity constraints. We show that we can slightly alter his analysis to incorporate the two inefficiencies discussed above, and we derive sufficient-statistics equations for the derivative of welfare with respect to the tuition subsidy, which we use to evaluate how climate change and imperfect altruism each impact the optimal subsidy. The estimated effects are large: in our baseline case, the optimal subsidy from (Lawson 2017) was about $8000 per year, and we find that the optimal subsidy increases by about $5000 per year in the presence of climate change, and decreases by over $7000 per year in the presence of imperfect parental altruism. The significant magnitude of these impacts are consistent with the intuition from (Hendren and Sprung-Keyser 2020) that policies that impact very young children are likely to have particularly large welfare impacts; in our case, the children are impacted before they are even born.

It is important to note that the goal of this paper is not to calculate a precise estimate of the optimal college tuition subsidy. Rather, our goal is to demonstrate that public economists can incorporate population externalities in their analyses of public policy, and that the impact of such externalities can be very significant. Accordingly, a final section of our paper shows how our approach can be generalized to other policies: under certain assumptions, the population externality can be expressed as the product of the effect of the program on fertility (which is program-specific) and a term that captures the magnitude of the distortion to fertility.

In our analysis, we assume a unitary family model for simplicity, in which a college tuition subsidy offered to the one unit of parent impacts the number of units of child produced. However, this is isomorphic to a setting in which each family consists of a father and a mother, if we assume that the government is constrained to set the same subsidy for males and females and that the responsiveness of enrollment to subsidies is the same for male and female students. As we discuss in section 3, these are conservative assumptions, as the impact on optimal subsidies would be larger (either for women alone, or for all students, depending on the situation) if the government could set different subsidies for men and women or if women responded more to subsidies than men.[7]

The rest of the paper proceeds as follows. Section 2 presents a baseline model of education and tuition subsidies based on (Lawson 2017), and solves for the optimal tuition subsidy in that setting, while section 3 shows how dramatically the optimal policy can be affected – in either direction – by population externalities. Section 4 then extends this idea to a more general setting, showing how population externalities can affect analyses of other social policies. Section 5 concludes the paper.

2. Optimal Policy in (Lawson 2017) Model of Education

To begin our analysis, we need to first present our baseline model of education and tuition subsidies in the absence of fertility considerations. What follows is a simplified version of the presentation from (Lawson 2017), using the same notation as in that paper for simplicity. We then summarize the optimal policy analysis and results from that paper, which serve as a baseline for our analysis of the impact of fertility subsidies on optimal policy.

2.1 (Lawson 2017) Model

The model consists of 12 periods of 4 years each, representing a working-age life of 48 years. In the first period, each individual i in a population of measure one chooses between attending college ($s_i = 1$) and working at wage w_{01} ($s_i = 0$); we abstract from gender and consider a unitary household model, but we will discuss issues of gender further in section 3. In periods $t = 2,\ldots,12$, the individual works at an exogenous wage w_{st} that depends on their first-period education choice s_i, where $w_{1t} > w_{0t}$ $\forall t$. Individuals also choose labour supply l_{si} that is constant across time for simplicity, receiving an income of $Y_{sti} = w_{st} l_{si}$ per period. The real interest and discount rates are both equal to r, with the discount factor denoted as $\beta \equiv \frac{1}{1+r}$, and exogenous productivity growth causes wages to grow at a rate of g per period.

[7] If the government could set different tuition subsidies for male and female students, then the estimated zero effect of husband's education on fertility in (Amin and Behrman 2014) would imply that the optimal subsidy for men would be that from section 2 of this paper, whereas the impact on optimal subsidies for women would be much larger than those estimated in section 3 since the effect of their education on fertility would be twice as large as we assume there. Meanwhile, if women responded more to tuition subsidies than men, as indicated by (Dynarski 2008) and (Card and Lemieux 2001), then fertility would respond even more to tuition subsidies than we estimate, imply larger effects of population externalities on optimal subsidies.

Workers receive per-period utility $v^s(c^s_{vi}, l_{si})$ from consumption c^s_{vi} and labour supply, while students' utility $u(c_{ui})$ varies only with consumption, and utility functions satisfy the usual properties: $u', v^s_c > 0$, $v^s_l < 0$, and $u'', v^s_{cc}, v^s_{ll} < 0$, where subscripts denote derivatives. An individual who chooses not to attend college will select labour supply l_{0i} and consumption c^0_{vi} per period, and receive lifetime utility $U_{0i} = \sum_{t=1}^{12} \beta^{t-1}\, v^0(c^0_{vi}, l_{0i})$. Individuals who choose to attend college will select post-schooling consumption c^1_{vi} and labour supply l_{1i}, as well as consumption c_{ui} while in college, receiving lifetime utility $U_{1i} + \eta_i$, where $U_{1i} = u(c_{ui}) + \sum_{t=2}^{12} \beta^{t-1}\, v^1(c^1_{vi}, l_{1i})$ and where η_i represents idiosyncratic utility from schooling and is the only source of heterogeneity in this simplified form of the model.

We define $R_x \equiv \sum_{t=x}^{12} \left(\frac{1}{1+r}\right)^{t-1}$ and $\gamma_x \equiv \sum_{t=x}^{12} \left(\frac{1+g}{1+r}\right)^{t-1}$ to simplify notation, and then we can write the individual's budget constraints, for $s_i = 0$ and $s_i = 1$ respectively:

$$R_1 c^0_{vi} = (1-\tau)\gamma_1 w_{01} l_{0i}$$

$$c_{ui} + R_2 c^1_{vi} = (b-e) + (1-\tau)\gamma_2 w_{11} l_{1i}$$

where e is the direct cost of college to the individual, τ is the marginal tax rate, and b is the government grant given to students, which incorporates all financial support provided to students by the government. Students may also face a liquidity constraint in the form of a debt limit A:

$$c_{ui} - (b-e) \leq A.$$

Therefore, the individual's maximization problem is to choose $\{s_i, c^0_{vi}, c^1_{vi}, c_{ui}, l_{0i}, l_{1i}\}$ to maximize $V_i = s_i(U_{1i} + \eta_i) + (1 - s_i)U_{0i}$:

$$\begin{aligned} V_i \;&= s_i[u(c_{ui}) + R_2 v^1(c^1_{vi}, l_{1i}) + \eta_i - \lambda_{1i}(c_{ui} + R_2 c^1_{vi} - (b-e) \\ &- (1-\tau)\gamma_2 w_{11} l_{1i}) - \mu_i(c_{ui} - (b-e) - A)] + (1-s_i)[R_1 v^0(c^0_{vi}, l_{0i}) \\ &- \lambda_{0i}(R_1 c^0_{vi} - (1-\tau)\gamma_1 w_{01} l_{0i})]. \end{aligned} \tag{1}$$

The government chooses b and τ subject to a budget constraint:

$$Sb + G = \tau[S\gamma_2 E(Y_{11i}|s_i = 1) + (1-S)\gamma_1 E(Y_{01i}|s_i = 0)] = \tau\bar{Y}$$

where $S = E(s_i)$ is the fraction of the population attending college, $\bar{Y}$ is mean total discounted lifetime income, and G is the exogenous discounted sum of other government spending over the 12 periods. Social welfare $V = E(V_i)$ is utilitarian with equal weights, and an envelope theorem result then simplifies our welfare analysis: any changes in individual choices of labour supply, consumption, and college enrollment when b changes have no first-order welfare impact, because those variables were chosen by the individual to maximize their utility. As a result, welfare can be written as $V(b, \tau(b))$, and the welfare gain from increasing b is:

$$\frac{dV}{db} = \frac{\partial V}{\partial b} + \frac{\partial V}{\partial \tau}\frac{d\tau}{db} = E\left(\frac{\partial V_i}{\partial b}\right) + E\left(\frac{\partial V_i}{\partial \tau}\right)\frac{d\tau}{db}. \tag{2}$$

This welfare derivative combines the direct effect of raising b with the indirect effect of changing τ to balance the government budget. To evaluate it, we use the first-order conditions of the individual's maximization problem. Since η_i is the only source of heterogeneity, consumption and labour supply choices do not vary by individual conditional on the choice of s_i:

$$\begin{aligned}
\frac{\partial V_i}{\partial b} &= s_i(\lambda_{1i} + \mu_i) = s_i u'(c_u) \\
\frac{\partial V_i}{\partial \tau} &= -s_i \lambda_{1i}\gamma_2 Y_{11i} - (1 - s_i)\lambda_{0i}\gamma_1 Y_{01i} \\
&= -s_i\gamma_2 Y_{11} v_c^1(c_v^1, l_1) - (1 - s_i)\gamma_1 Y_{01} v_c^0(c_v^0, l_0) \\
\frac{d\tau}{db} &= \frac{S}{\bar{Y}}\left[1 + \varepsilon_{Sb} - \left(1 + \frac{G}{Sb}\right)\varepsilon_{\bar{Y}b}\right]
\end{aligned} \tag{3}$$

where ε_{Sb} is the elasticity of college enrollment with respect to student grants b, and $\varepsilon_{\bar{Y}b}$ is the elasticity of average income (accounting for foregone earnings) with respect to b.

In the expression for $\frac{\partial V_i}{\partial \tau}$, we make the conservative assumption that $S\gamma_2 Y_{11} v_c^1(c_v^1, l_1) + (1 - S)\gamma_1 Y_{01} v_c^0(c_v^0, l_0) \simeq \bar{Y} v_c^0(c_v^0, l_0)$, and we also normalize the welfare gain into a dollar amount, by defining $\frac{dW}{db} \equiv \frac{\frac{dV}{db}}{v_c^0(c_v^0, l_0)}$ as the welfare gain in terms of an equivalent amount of consumption among non-graduates. The welfare derivative thus becomes:

$$\frac{dW}{db} \simeq S\left[\frac{u'(c_u) - v_c^0(c_v^0, l_0)}{v_c^0(c_v^0, l_0)} - \varepsilon_{Sb} + \left(1 + \frac{G}{Sb}\right)\varepsilon_{\bar{Y}b}\right]. \tag{4}$$

Finally, we decompose the ratio of marginal utilities into two empirically observable quantities, which can be called liquidity and substitution effects. An individual attends college if the idiosyncratic utility from schooling exceeds a critical value η^*:

$$\eta^* = R_1 v^0(c_v^0, l_0) - u(c_u) - R_2 v^1(c_v^1, l_1).$$

We assume that η_i follows some continuously differentiable distribution $F(\eta)$ with a density given by $f(\eta)$. As a result, $S = 1 - F[R_1 v^0(c_v^0, l_0) - u(c_u) - R_2 v^1(c_v^1, l_1)]$, and therefore:

$$\frac{\partial S}{\partial b} = f(\eta^*)u'(c_u)$$

$$\frac{\partial S}{\partial a_1} = f(\eta^*)[u'(c_u) - v_c^0(c_v^0, l_0)]$$

where a_1 is a lump-sum of cash in the first period, representing a change in initial assets. Thus, we can rewrite (4)(2) as:

$$\frac{dW}{db} \simeq S\left[L - \varepsilon_{Sb} + \left(1 + \frac{G}{Sb}\right)\varepsilon_{\bar{Y}b}\right] \tag{5}$$

where $L = \frac{\frac{\partial S}{\partial a_1}}{\frac{\partial S}{\partial b} - \frac{\partial S}{\partial a_1}}$. The $\frac{\partial S}{\partial a_1}$ in the numerator of L is the liquidity effect, whereas the $\frac{\partial S}{\partial b} - \frac{\partial S}{\partial a_1}$ in the denominator is the substitution effect, as it represents the effect on enrollment of changing relative prices without transferring income to students. A higher value of L thus indicates more severe liquidity constraints among students.

2.2 Optimal Tuition Subsidies

As in (Lawson 2017), we can now use estimates of each of the terms in (5) – the sufficient statistics for welfare analysis – to calculate an estimated value of $\frac{dW}{db}$, and we can also perform statistical extrapolations of these quantities, modelling how their values change as b changes to find the optimal tuition subsidy b. The list below gives a brief summary of the values chosen for the sufficient statistics in Table 1, which are chosen to represent the United States in 2007; hats represent baseline values, and further details can be found in (Lawson 2017).

- To estimate the baseline b, we use data on receipt of federal and state grants, loans, and work-study in 2007-08 from (Wei et al. 2009), and an adjustment formula from (Epple, Romano, and Sieg 2006); financial support was about \$1690 per student, and lacking data on other forms of government assistance, we round this up to \$2000. As we will denote monetary amounts in thousands of dollars per year, we have $\hat{b} = 2$.
- (Deming and Dynarski 2009) find a general consensus that a \$1000 increase in price of college leads to a 4 percentage point decline in attendance, which implies an elasticity of $\varepsilon_{Sb} \simeq 0.2$. As a lower bound, we use the fact that (Dynarski 2008) estimates that \$2500 of financial aid leads to a 4 percentage point increase in degree completion from a base of 27% to motivate a value of 0.1.
- The estimated college enrollment rate of 18-24-year-olds in 2007 was $\hat{S} = 0.388$, according to. Assuming a constant elasticity of enrollment with respect to grants, this implies that $S = \phi b^{\varepsilon_{Sb}}$, where $\phi = \frac{\hat{S}}{\hat{b}^{\varepsilon_{Sb}}}$.
- Our preferred estimate of L is zero, as numerous papers argue that income has no causal effect on enrollment. Alternatively, results in (Acemoglu and Pischke 2001) imply that a \$1000 increase in family income increases enrollment by 0.21% points, so we also consider $\widehat{\frac{\partial S}{\partial a_1}} = 0.0021$, which implies $\hat{L} = 0.057$ or 0.121 depending on the value of ε_{Sb}. As in (Lawson 2017), we model a decline in L with S according to $L = \max\{\frac{\hat{L}(0.16-(S-\hat{S}))}{0.16+\hat{L}(S-\hat{S})}, 0\}$.
- We assume that the interest and discount rates are 3% per year, which implies $r = 0.12$ and $\beta = \frac{1}{1.12} \simeq 0.893$. We assume wage growth of $g = 0.04$, since the average real growth rate in the SSA average net compensation series was 1% over 1991–2008.
- We assume that each year of schooling increases earnings by 8%, and that the elasticity of taxable income is 0.4, as found by (Gruber and Saez 2002); see (Lawson 2017) for the resulting algebraic expression for $\varepsilon_{\bar{Y}b}$, with baseline values of 0.0063 or 0.0142 depending on the value of ε_{Sb}.
- The baseline tax rate is $\tau = 0.23$ (incorporating a 15% federal tax, a 5% state tax, and 3% for the Medicare tax), and the CPS 2008 Annual Social and Economic Supplement gives us $Y_{01} = 34$ for high school workers, meaning that $Y_{11} = 34(1.08)^4 = 46.26$. These estimates imply $G = 68.606$ and $\frac{\hat{G}}{Sb} = 88.410$, and thus $\frac{G}{Sb} = 88.41\frac{\hat{S}\hat{b}}{Sb}$.

Table 1. Baseline Values of Sufficient Statistics

Statistic	*Definition*	*Value*
$\hat{S}$	enrollment rate	0.388
$\hat{b}$	per-year student grant	2
ε_{Sb}	elasticity of enrollment w.r.t. b	{0.1,0.2}
$\widehat{\frac{\partial S}{\partial a_1}}$	effect of income on enrollment	{0,0.0021}
r	interest and discount rate per period	0.12
g	wage growth per period	0.04
$\hat{\varepsilon}_{\bar{Y}b}$	elasticity of mean income w.r.t. b	{0.0063,0.0142}
$\widehat{\frac{G}{Sb}}$	ratio of exogenous spending to grant spending	88.410

The optimal policy results, as estimated in (Lawson 2017), can be found in Table 2. Panel A presents an estimate of the welfare derivative at the baseline $b = 2$, in dollar terms relative to the annual amount within a period, so it is the present-value equivalent of the welfare derivative expressed per year over the 4 years of the first period. Panel B contains the estimated optimal grants, and panel C lists the estimated net welfare gains from moving to the optimum, which are calculated by numerically integrating $\frac{dW}{db}$ from $b = 2$ to the optimum; the welfare gain is then expressed as the dollar amount of an equivalent one-year per-person consumption increase, as well as (in brackets) a percentage of the initial size of the student grant program.

The results from the (Lawson 2017) analysis are dramatic: student grants should be increased by at least \$3800 per year, and by over \$6000 in our preferred case of $\varepsilon_{Sb} = 0.2$ and $\widehat{\frac{\partial S}{\partial a_1}} = 0$. The welfare gains are also large, at about 41 cents per dollar spent at baseline in our preferred case, and with the optimal policy generating welfare improvements that are equivalent to as much as 0.17% of GDP at the economy-wide level. In our preferred case, enrollment increases from 38.8% to 51.3% at the optimal $b = 8.093$.

Table 2. Optimal Policy Results using (5)

	$\widehat{\frac{\partial S}{\partial a_1}}$	
ε_{Sb}	0	0.0021
	A. Estimate of $\frac{dW}{db}$ at $b = 2$	
0.1	0.1811	0.2282
0.2	0.4148	0.4370
	B. Optimal Student Grants	
0.1	\$5843	\$8371
0.2	\$8093	\$8355
	C. Welfare Gains from Moving to Optimum	
0.1	\$947 (30.5%)	\$1750 (56.4%)
0.2	\$3138 (101.1%)	\$3471 (111.8%)

Notes: Panel A presents the one-period per-year increase in welfare, expressed in dollars of consumption from a per-year one dollar increase in b. *Panel C expresses welfare gains as a one-time lump-sum increase in consumption, and as a percentage of baseline spending on student grants. This format is used throughout all subsequent tables of this form.*

3. Optimal Policy with Population Externalities

The model to this point has simply been a slightly simplified version of the baseline model from (Lawson 2017). We can now build upon this baseline by introducing a new population dimension to our model, in the form of a fertility decision by the model's individuals. Specifically, we assume that, in period 3 of the model, each individual i chooses how many children n_{si} to have, where the decision may depend on their s_i. The third period corresponds to ages 26–30, and (OECD Family Database 2019) indicates that a mother's average age at childbirth is about 29 in the United States; we avoid issues of discreteness by allowing n_{si} to be a continuous variable that represents expected children per individual.

We abstract from gender by considering a unitary family model in which the individual represents the mother-father unit, and we use the estimate from (Amin and Behrman 2014) that the causal impact of university education relative to high school is -0.678 children for women and roughly zero for men, which in a unitary-household model implies that the negative effect of university on fertility is -0.339 units of child per household. It is important to note that this is isomorphic to a setting in which families consists of a separate father and mother, if we assume that the government must set the same subsidy for men and women and that the responsiveness of enrollment to subsidies is the same for men and women.

It is also important to note that the latter are conservative assumptions that are likely to lead us to an underestimate of the impact of population externalities (either for all students or for women alone). If the government can set different tuition subsidies for men and women, then the finding in (Amin and Behrman 2014) that the effect of the husband's education on fertility is zero means that the father's optimal subsidy would be given by section 2 of this paper (with no fertility impacts), whereas the mother's optimal subsidy would be given by the analysis to come but with a doubled effect of education on fertility (-0.678 rather than -0.339). In that case, the effect of population externalities on optimal tuition subsidies, at least for women, would be even larger than estimated in this paper.

Alternatively, we could apply the finding in (Dynarski 2008) that the overall effect of financial aid on graduation is higher for women,[8] which would imply that fertility responds even more to tuition subsidies, since those subsidies will impact women – whose education matters for fertility – more than the average. This would again imply even larger population externality effects than what we estimate in the current section of the paper.

For simplicity, we assume that the household's utility from children is additively separable from the rest of the utility function (1), taking the form of net private utility $q_s(n_{si})$, where the utility function varies according to whether an individual went to university or not, which will permit different fertility levels to be chosen by college graduates and non-graduates. Conditional on their choice of $s_i = \{0,1\}$, the individual will choose n_{si} such that $q'_s(n_{si}) = 0$; however, we will consider below the possibility that q_s is not equal to the social value of n_{si} children, so that the privately-optimal level of fertility may not be socially optimal.

In particular, we will now model – one at a time – the two possible population externalities that we consider in this paper: environmental externalities from greenhouse gases, and imperfect parental altruism.

3.1 Environmental Externalities from Population

We now introduce environmental externalities from population: we will assume that, while $q_s(n_{si})$ represents the private utility of individual i from their children,

[8] (Card and Lemieux 2001) finds similar evidence of a greater responsiveness of female enrollment to tuition fees. On the other hand, while very few studies estimate effects of price on enrollment separately for men and for women, studies that consider only men – such as (Cameron and Heckman 2001) – find estimates of similar magnitude to the rest of the literature.

those children will also contribute to a public bad in the form of climate change. For simplicity, we assume this is the only effect of education on greenhouse gas emissions; that is, we assume that there is no direct effect of education on emissions (or that a carbon tax has been set optimally to offset this margin).[9] It is possible that a direct effect of education on emissions could provide an alternative mechanism altering the optimal subsidy policy, but we abstract from this mechanism in our paper to focus on the fertility channel.

As a result of these assumptions, we can write a new social welfare function that extends (1) to this setting. Since the function q_s does not vary across individuals, there are only 2 possible levels of fertility in our model (n_0 and n_1), and so we can simplify our notation from n_{si} to n_s, giving us:

$$\hat{V}_i = V_i + s_i q_1(n_1) + (1 - s_i) q_0(n_0) - e(\bar{n}) \tag{6}$$

where V_i is identical to the original equation (1), and it is followed by 3 new terms. The first two represent the private utility from fertility (which is also valued by the social planner), whereas the final term represents the disutility experienced by everyone from the pollution generated by children, where $\bar{n} = Sn_1 + (1 - S)n_0$ is the average number of children in the population.

Due to our simplifying assumption of additive separability of utility from children, the subsequent welfare analysis is unchanged from that in section 2 except for two things. First of all, the critical value η^* has to be replaced with a critical value of $\eta_i + q_1(n_1) - q_0(n_0)$, but this has no impact on the results, as it is simply a redefining of variables which are not among the sufficient statistics. Second, and more importantly, the welfare derivative in (2) has an extra term:

$$\frac{d\hat{V}}{db} = \frac{\partial V}{\partial b} + \frac{\partial V}{\partial \tau}\frac{d\tau}{db} + \frac{\partial \hat{V}}{\partial \bar{n}}\frac{d\bar{n}}{db}$$

because each n_{si} is the value that is privately optimal to the individual, but not the socially optimal value given the externality that operates through $\bar{n}$, which no individual is large enough to affect on their own.

[9] (Bruderer Enzler and Diekmann 2015) provide a summary of the literature that studies the impact of individual characteristics on individual greenhouse gas emissions, and finds mixed results; a more recent paper by (Moser and Kleinhückelkotten 2018) finds no significant effect of education on emissions. However, the consensus in the literature is that higher income raises emissions, so the overall effect of increased education is probably an increase in emissions.

Therefore, since $\bar{n}$ changes with b only because S changes with b, the welfare derivative in dollar terms can be written as:

$$\frac{d\widehat{W}}{db} = \frac{dW}{db} - \frac{e'(\bar{n})}{v_c^0(c_v^0, l_0)} \frac{d\bar{n}}{dS} \frac{dS}{db} \tag{7}$$

and to evaluate this derivative, we simply need to calculate the final term and subtract it from the $\frac{dW}{db}$ calculated in section 2. We already know that $\frac{d\bar{n}}{dS} = -0.339$, and that $\frac{dS}{db} = \frac{S}{b}\varepsilon_{Sb}$, so the main challenge is calculating $\frac{e\prime(\bar{n})}{v_c^0(c_v^0, l_0)}$, or the social environmental disutility caused by population in dollar terms.

To calculate this term, we need estimates of the emissions per person per year, the (discounted) number of years per person, and the social cost of each unit of emissions. For the first of these quantities, we use the estimate from (Climate Watch 2019) that carbon-dioxide-equivalent greenhouse gas emissions are about 20 tonnes per person in the United States; we assume that this average is also the marginal amount, as we are not aware of any estimates of the emissions impact of adding a marginal person to the population.

We assume that these 20 tonnes per year will be emitted by our marginal individual over a life expectancy of 20 periods, starting from period 4; however, we need to discount the costs of emissions occurring in the future. We continue to use a discount rate of 3% per year, but estimates from (EPA 2017) indicate that the social cost of carbon emissions also increases in real terms over time, at a rate of about 2% per year.[10] With a resulting net discount rate of 1% per year or 4% per period, this implies that the 20 periods of life are the discounted present value equivalent of 12.565 periods.

Finally, we assume a social cost of carbon of $100 per tonne. This value is a compromise between actual values used in the real world and recent estimates of the actual social cost from the scientific literature. For example, in 2015, the (EPA 2017) assumed a social cost of carbon of $36 in 2007 dollars,[11] whereas (Ricke et al. 2018) estimate a global social cost of carbon of $417 per tonne of carbon dioxide.

Therefore, a child born in our model will generate a social cost that is equivalent to 12.565 × 20 × 100 = $25130 per year over a 4-year period from their contribution

[10] (EPA 2017) explains that "estimates of the social cost of these greenhouse gases increase over time because future emissions are expected to produce larger incremental damages as physical and economic systems become more stressed in response to greater climatic change, and because GDP is growing over time and many damage categories are modeled as proportional to gross GDP."

[11] An exception to the rule that the values used in the real world are relatively low is given by Sweden, where the current carbon price is 110 Euros per tonne; see (Government Offices of Sweden 2020).

to climate change, and so $\frac{e\prime(\bar{n})}{v_c^0(c_v^0,l_0)} = 25.13$. This means that the term that we add to the welfare derivative $\frac{dW}{db}$ is $8.52\frac{S}{b}\varepsilon_{Sb}$, and we can produce new estimates of the optimal policy results from Table 2 in the current model, which can be found in Table 3.

The welfare implications of the population externality from climate change are highly significant: the welfare derivatives nearly double, and the optimal grants increase by at least $3500, with an increase of $5000 per year in our preferred case of $\varepsilon_{Sb} = 0.2$ and $\frac{\widehat{\partial S}}{\partial a_1} = 0$. The overall welfare gains are also 2 to 3 times as large, and college enrollment S increases to 56.5% at the optimal policy in our preferred scenario. In additional results that are available upon request, we can produce estimates from simulations of the calibrated model instead of statistical extrapolation; as in (Lawson 2017), those results are similar to the ones in Table 3.[12]

We can also perform a sensitivity analysis with a social cost of carbon of $417 as in (Ricke et al. 2018), and the impact of population externalities is extremely large in that case: in the preferred case of $\varepsilon_{Sb} = 0.2$ and $\frac{\widehat{\partial S}}{\partial a_1} = 0$, the welfare derivative increases to 1.7932, and the optimal grant is $29106, which is approximately equal to the average worker's after tax income at baseline. The welfare gain becomes $33279, equivalent to a 1.6% increase in GDP each year, and college enrollment S increases to 66.3% at the optimum. These results can most plausibly be interpreted as saying that, if the social cost of carbon is really $417, we need to do a lot to reduce climate change, including a significant reduction to fertility – and we probably need to go well beyond education policy in doing so.

In any case, we can conclude that population externalities from climate change have a large impact on optimal tuition subsidy policy: if we are concerned about the contribution of population to climate change, but we lack efficient direct policy tools to affect fertility, very large subsidies to college tuition would be justified if that would lead to reductions in fertility.

[12] The welfare derivatives are slightly smaller in the simulations, but when $\varepsilon_{Sb} = 0.2$, the optimal grant is about $1000 higher and the overall welfare gain is about 50% higher. The optimal b and the welfare gain are also significantly larger when $\varepsilon_{Sb} = 0.1$ and $\frac{\widehat{\partial S}}{\partial a_1} = 0$.

Table 3. Optimal Policy Results using (7)

	$\frac{\widehat{\partial S}}{\partial a_1}$	
ε_{Sb}	0	0.0021
	A. Estimate of $\frac{d\widehat{W}}{db}$ at $b = 2$	
0.1	0.3464	0.3935
0.2	0.7454	0.7676
	B. Optimal Student Grants	
0.1	\$9380	\$12429
0.2	\$13072	\$12879
	C. Welfare Gains from Moving to Optimum	
0.1	\$2826 (91.0%)	\$4115 (132.6%)
0.2	\$8338 (268.6%)	\$8694 (280.1%)

3.2 Imperfect Parental Altruism

An alternative potential reason for fertility deviating from the social optimum can be found in the field of population ethics: we assume that parents receive private utility $q_s(n_{si})$ from their children, and we now allow for the possibility that the real social utility $\tilde{q}_s(n_{si})$ generated by the children is larger than q_s. That is, the parent might not be perfectly altruistic and might underweight the utility of their children relative to what a utilitarian social planner would do if that planner wanted to maximize the discounted stream of all present and future utilities in the world.[13]

To model this possibility, we will explicitly specify both the parental utility function $q_s(n_{si})$ and the social utility function $\tilde{q}_s(n_{si})$. We assume that the utility function from children consists of two terms: a term that is linear in n_{si}, representing the utility per person that will be experienced by children, and a cost term that is increasing and convex in n_{si}, representing the monetary and/or effort cost for the parent to raise n_{si} children. We will then assume that the parent underweights the first term, underappreciating the utility that will be experienced by their children in the future.

[13] Alternative mechanisms that could generate inefficiently low fertility might include market-size effects that cause larger populations to discover a larger number of positive innovations, though the functional forms might be different in that case.

Before presenting the utility functions, we have to address an important question: why do more-educated individuals have fewer children? In the previous subsection, we did not explicitly model the parental utility functions, as envelope conditions easily eliminated them from the welfare derivative, but if we are to model $q_s(n_{si})$, we have to model the mechanism that allows education to affect fertility. There are two possibilities: more-educated individuals' costs are higher, or their perceived benefits in terms of future child utility are lower. The second approach seems problematic, as it implies that more-educated individuals are less altruistic than they would be if they had less education (which could make individuals reluctant to attend college as that would reduce their altruistic utility). Instead, we choose the first approach and assume that more-educated individuals have fewer children because higher education raises their costs of child-bearing; this seems plausible, if more-educated individuals have more valuable uses for their time. Therefore, we assume that the net parental utility function is:

$$q_s(n_{si}) = \alpha \hat{u}_2 n_{si} - \frac{\theta_s}{\chi} n_{si}^{\chi}$$

whereas the social utility function from children is:

$$\tilde{q}_s(n_{si}) = \hat{u}_2 n_{si} - \frac{\theta_s}{\chi} n_{si}^{\chi}$$

where $\alpha < 1$ represents the degree of imperfect altruism and $\chi > 1$ represents the degree of convexity in the cost function. $\hat{u}_2 \equiv u_2 - \bar{u}_2$ represents the expected utility u_2 of the child relative to a critical value $\bar{u}_2$;[14] we assume that the child's expected utility does not vary with the education decision of the parent, to keep the algebra simple. θ_s is the parameter that determines the level of the cost function, and we assume that $\theta_1 > \theta_0$ as discussed above. As a result, the number of children chosen by individuals depends only on their education decision, with $n_s = \left(\frac{\alpha \hat{u}_2}{\theta_s}\right)^{\frac{1}{\chi - 1}}$, and so $n_1 < n_0$ as expected.

At the private optimum, the utility experienced by the parent from their children is then given by:

$$q_s(n_s) = \left(\frac{\chi - 1}{\chi}\right) \theta_s \left(\frac{\alpha \hat{u}_2}{\theta_s}\right)^{\frac{\chi}{\chi - 1}}$$

[14] The critical value is necessary because a value of zero in our utility functions does not necessarily have an obvious economic interpretation.

whereas the social utility function at that same value of n_s is:

$$\tilde{q}_s(n_s) = \left(\frac{\chi - \alpha}{\alpha\chi}\right)\theta_s\left(\frac{\alpha\hat{u}_2}{\theta_s}\right)^{\frac{\chi}{\chi-1}}$$

The planner has no way of affecting the value of n_s chosen by the parent conditional on s_i, but if they want to encourage fertility, they can try to discourage individuals from going to college. To be precise about this intuition, we can solve for the welfare derivative, starting from the welfare function:

$$\hat{V}_i = V_i + s_i\tilde{q}_1(n_1) + (1 - s_i)\tilde{q}_0(n_0) \tag{8}$$

where as before, the critical value η^* will need to be replaced with a critical value of $\eta_i + q_1(n_1) - q_0(n_0)$, which simply shifts the mean of the distribution of η.

It is now simplest to write the extra term of the welfare derivative in terms of S, because s_i is not chosen efficiently due to the difference between the social utility function $\tilde{q}_s$ and the private utility function q_s:

$$\frac{d\hat{V}}{db} = \frac{\partial V}{\partial b} + \frac{\partial V}{\partial \tau}\frac{d\tau}{db} + \frac{\partial \hat{V}}{\partial S}\frac{dS}{db}$$

and for the partial derivative with respect to S, we have:

$$\frac{\partial \hat{V}}{\partial S} = \frac{\partial V}{\partial S} + \tilde{q}_1(n_1) - \tilde{q}_0(n_0)$$

whereas the individual's optimization condition gives us:

$$\frac{\partial V}{\partial S} + q_1(n_1) - q_0(n_0) = 0.$$

Therefore our partial derivative with respect to S becomes:

$$\begin{aligned}\frac{\partial \hat{V}}{\partial S} &= (\tilde{q}_1(n_1) - q_1(n_1)) - (\tilde{q}_0(n_0) - q_0(n_0)) \\ &= \left(\frac{1-\alpha}{\alpha}\right)(\alpha\hat{u}_2)^{\frac{\chi}{\chi-1}}\left[\theta_1^{\frac{-1}{\chi-1}} - \theta_0^{\frac{-1}{\chi-1}}\right]\end{aligned}$$

and this can even be further simplified if we can observe n_1 and n_0, because we can use the individual's solution for $n_s = \left(\frac{\alpha\hat{u}_2}{\theta_s}\right)^{\frac{1}{\chi-1}}$, and therefore we have:

$$\frac{\partial \hat{V}}{\partial S} = (1-\alpha)\hat{u}_2(n_1 - n_0).$$

Therefore, our welfare derivative now becomes:

$$\frac{d\widehat{W}}{db} = \frac{dW}{db} - \frac{\hat{u}_2}{v_c^0(c_v^0, l_0)}(1-\alpha)(n_0 - n_1)\frac{dS}{db}. \tag{9}$$

We already know that $n_0 - n_1 = 0.339$, and that $\frac{dS}{db} = \frac{S}{b}\varepsilon_{Sb}$, so the main challenges are calculating $\frac{\hat{u}_2}{v_c^0(c_v^0,l_0)}$ (the social utility generated by a child in dollar terms), and choosing a value of α.

To estimate $\frac{\hat{u}_2}{v_c^0(c_v^0,l_0)}$, we simulate the calibrated model from (Lawson 2017) for the childrens' generation, using the parameters from that paper and assuming that wages (and the debt limit A) continue to grow at $g = 0.04$ per period for the child. We also increase the mean of η_i to keep the percentage of children who go to university at $\hat{S} = 0.388$, and calculate the expected lifetime utility of the child, which is then discounted back to the start of the model to give us a total of $u_2 = 11.9821$ in utils. The critical level $\bar{u}_2$ is somewhat arbitrary, but to be very conservative, we assume that it is the level of utility obtained by an uneducated child who works full-time at the 2007 federal minimum of \$5.85. Such a worker obtains consumption of 11.7 with $l_{0i} = 1$, which generates a total discounted present value utility of $\bar{u}_2 = 3.4298$.

Finally, at the baseline values in (Lawson 2017), the marginal utility from consumption for an uneducated worker is $v_c^0(c_v^0, l_0) = 0.045$, which means that $\frac{\hat{u}_2}{v_c^0(c_v^0,l_0)} = 190.1049$. Our conservative estimate, therefore, is that the discounted present social value of the utility of a child is about \$190 thousand dollars per year over 4 years, or about \$760 thousand dollars in total. Therefore, the term we subtract from the welfare derivative $\frac{dW}{db}$ is $64.45(1-\alpha)\frac{S}{b}\varepsilon_{Sb}$, and we will present results for a value of $\alpha = 0.8$;[15] the optimal policy results can be found in Table 4.

[15] This is still a fairly high value for α, but for lower values the optimal grant drops to zero, which is an artificial lower bound caused by the parametric assumptions that we make, including the assumption that $S = 0$ if $b = 0$.

Table 4. Optimal Policy Results using (9)

	$\frac{\widehat{\partial S}}{\partial a_1}$	
ε_{Sb}	0	0.0021
	A. Estimate of $\frac{d\widehat{W}}{db}$ at $b = 2$	
0.1	-0.0709	-0.0242
0.2	-0.0853	-0.0633
	B. Optimal Student Grants	
0.1	\$536	\$1169
0.2	\$822	\$1021
	C. Welfare Gains from Moving to Optimum	
0.1	\$285 (9.2%)	\$46 (1.5%)
0.2	\$248 (8.0%)	\$145 (4.7%)

The results are dramatically different from those in the previous subsection on environmental externalities: the baseline welfare derivatives turn negative, the optimal grants decrease by about \$5300 to \$7300 per year, and the total welfare gains are now small because the optimal grants are not too far from $b = 2$. This is despite the fact that we made several conservative assumptions in the calibration, including the assumption that the critical level of utility is that of a full-time minimum wage worker – which implies that anyone worse off would have, from the perspective of the planner, a life that is not worth living – and our assumption that parents only underestimate their child's contribution to global welfare by 20%. To make any of these assumptions less conservative, we would need to use a calibration-and-simulation approach, to model S more flexibly. In principle, however, we could generate almost any magnitude of negative b – that is, of education taxes – with empirically plausible assumptions on the parameters.

As before, we can conclude that population externalities have a large impact on optimal tuition subsidy policy in this case: if we are concerned about the population ethics consequences of imperfect altruism, but we lack efficient direct policy tools to affect fertility, any economic argument for increased subsidies to college tuition vanishes if such subsidies would lead to reductions in fertility.

4. A General Toolbox for Welfare Analysis

The analysis of this paper has focussed so far on the specific setting of college tuition subsidies, because this is a policy that can have straightforward and intuitive effects on fertility decisions. However, a variety of other social and economic policies could also have important impacts on fertility: obviously there are direct fertility incentives from "baby bonus" policies such as that studied by (Milligan 2005), but also maternal/parental leave and benefit programs (see (Hyatt and Milne 1991), (Phipps 2000), and (Cannonier 2014) for analyses of their effects on fertility), and there are a few studies of the impact of programs such as social assistance on fertility (Robins and Fronstin 1996). On a related note, (Currie and Schwandt 2014) find that higher unemployment reduces conceptions, which suggests that a policy like unemployment insurance that affects the duration and severity of experiences of unemployment could also have an impact on fertility.

Of course, if fertility is chosen in a socially-optimal way, then none of these potential interactions between public policy and fertility have first-order welfare implycations. However, if environmental or population ethics externalities exist, then any number of social policies that impact fertility could be affected by population externalities.

As a result, we will now demonstrate how the analysis from the previous section of the paper can be generalized to a broader range of policies. To do so, let us consider a version of the general model presented in section 3 of (Chetty 2009). We will focus on a static representative agent model, but the analysis easily generalizes to heterogeneous agents and multi-period problems, as mentioned by (Chetty 2009). To make the intuition of the basic model as easy to understand as possible, we begin with a version without any population externalities; in other words, there are no future generations to consider (or at least their utility is unaffected by choices made by the current generation).

We assume that the representative agent makes a vector of J choices denoted by $x = \{x_1, \ldots, x_J\}$, receiving utility $U(x)$. The vector x can include choices such as the binary decision to attend university, but it also includes consumption and labour supply in different states, such as graduate, non-graduate, and student, along with potentially a wide variety of other choice variables.

The government, meanwhile, operates a variety of programs and taxes that act upon various choices made by the agent, but as in (Chetty 2009) we assume that the government wants to evaluate one particular program, which we assume is a monetary incentive acting upon choice 1. The government places this monetary incentive p (which could be a subsidy to education, or any other one-parameter policy program) on x_1, and finances it with a tax t on x_J, where a government budget constraint (which may include other exogenous government spending as in the education model above) defines the budget-balancing values of t given p and the agent's choices.

The agent also faces a set of $M < J$ constraints on their choices, given by $\{G_1(x,p,t) = 0, \ldots, G_M(x,p,t) = 0\}$, and so the representative agent solves:

$$\max_x U(x) \;\; s.t. \;\; G_1(x,p,t) = 0, \ldots, G_M(x,p,t) = 0. \tag{10}$$

Therefore, the equation for social welfare can be written as:

$$W(p,t) = \max_x \left\{ U(x) + \sum_{m=1}^{M} \lambda_m G_m(x,p,t) \right\}. \tag{11}$$

When we take the derivative of W with respect to p, we can use the envelope condition resulting from individual maximization to conclude that $\frac{\partial W}{\partial x} = 0$, and therefore the welfare derivative is:

$$\frac{dW}{dp} = \sum_{m=1}^{M} \lambda_m \left[\frac{\partial G_m}{\partial p} + \frac{\partial G_m}{\partial t} \frac{dt}{dp} \right] \tag{12}$$

where $\frac{dt}{dp}$ comes from the total derivative of the government budget constraint. We then use the individual maximization results:

$$U_j(x) = -\sum_{m=1}^{M} \lambda_m \frac{\partial G_m}{\partial x_j} \;\; \forall j$$

where $U_j(x)$ is the derivative of U with respect to x_j. Next, we add an assumption that (Chetty 2009) calls Assumption 1, which is that p affects the constraints in the same way as the good (x_1) on which it is placed, and the tax affects the constraints in the same way as the good (x_J) on which it is levied. This implies that there exist functions $k_p(x,p,t)$ and $k_t(x,p,t)$ such that:

$$\frac{\partial G_m}{\partial p} = -k_p(x,p,t) \frac{\partial G_m}{\partial x_1} \;\; \forall m$$
$$\frac{\partial G_m}{\partial t} = k_t(x,p,t) \frac{\partial G_m}{\partial x_J} \;\; \forall m.$$

Using these results and assumptions, we can write the welfare derivative as follows:

$$\begin{aligned}\frac{dW}{dp} &= \sum_{m=1}^{M}\lambda_m\left[-k_p(x,p,t)\frac{\partial G_m}{\partial x_1}+k_t(x,p,t)\frac{\partial G_m}{\partial x_J}\frac{dt}{dp}\right]\\ &= -k_p(x,p,t)\sum_{m=1}^{M}\lambda_m\frac{\partial G_m}{\partial x_1}+k_t(x,p,t)\frac{dt}{dp}\sum_{m=1}^{M}\lambda_m\frac{\partial G_m}{\partial x_J}\\ &= k_pU_1(x)-k_t\frac{dt}{dp}U_J(x).\end{aligned}$$

To be able to implement this welfare derivative empirically, we need to be able to evaluate each of these terms. The model itself generally provides functional forms for k_p, k_t, and the derivative of the government budget constraint $\frac{dt}{dp}$, whereas the marginal utility terms require a further step which varies depending on the model: they need to be written in terms of empirically observable quantities, using direct estimates of marginal utility or using the structure of the model to make a substitution. For example, in the education model from section 2 of this paper, $\frac{d\tau}{db}$ takes the place of $\frac{dt}{dp}$, $u'(c_u)$ and v_c^* are the heterogeneous-agent versions of $U_1(x)$ and $U_J(x)$, and $k_p = S$ and $k_t = \bar{Y}$.

We now switch our focus to population externalities; in particular, we will suppose that one of the x_j choices is fertility, and to avoid confusion we will call the fertility decision x_f. As we have stated throughout the paper, if x_f is chosen optimally by the agent, then the analysis above is completely unchanged. However, if the agent does not properly optimize x_f due to an externality, the analysis above will be incomplete, because we have failed to account for the possibility that changing p will affect the agent's choice of x_f.

To keep the analysis simple, let us assume that there is an external effect (which could be positive or negative) of x_f that is not accounted for by the individual; that is, the individual continues to solve the individual optimization problem (10), but the welfare function takes the following form:

$$\widehat{W}(p,t)=\max_x\left\{U(x)+\sum_{m=1}^{M}\lambda_mG_m(x,p,t)\right\}+e(x_f)$$

where $e(x_f)$ is the externality caused by fertility. It is important to note that, to keep the analysis simple, we assume additive separability, and we assume that this externality is only a function of x_f, with no interactions with the other x_j. That is, we assume that the externality is a function of the number of children born, but not of any characteristics of those children that might be a function of other choices represented by x_j. The functional form of $e(x_f)$ could depend on exogenous values of utility per child, or even on a critical value of utility in a setting of critical-level utilitarianism, but $e(x_f)$ cannot depend on an endogenous measure of utility per child. If we wanted to permit interactions between x_f and other x_j, we would need to consider the impact of p on those other choices as well, and we will return to a brief analysis of how this would affect our results later in this section.

The welfare derivative now takes the following form:

$$\frac{d\widehat{W}}{dp} = \sum_{m=1}^{M} \lambda_m \left[\frac{\partial G_m}{\partial p} + \frac{\partial G_m}{\partial t}\frac{dt}{dp}\right] + e'(x_f)\frac{dx_f}{dp}$$

so we simply need to add the extra term representing the effect of p on the externality $e(x_f)$. All of the subsequent steps above for $\frac{dW}{dp}$ are unchanged, as the individual maximization equations and the consequences of the assumption from (Chetty 2009) are unaffected by the introduction of a population externality. Therefore, the final welfare derivative is:

$$\frac{dW}{dp} = k_p U_1(x) - k_t \frac{dt}{dp} U_J(x) + e'(x_f)\frac{dx_f}{dp}$$

We find, therefore, that it is relatively simple to incorporate population externalities into a sufficient-statistic welfare analysis of a social program: we simply need to add an extra term that captures the effect of the program on fertility multiplied by the marginal effect of such a fertility impact on social welfare.

The effect of the program on fertility is given by $\frac{dx_f}{dp}$, which is a quantity that needs to be evaluated empirically; in our analysis, we use a value of $-0.339\frac{S}{b}\varepsilon_{Sb}$, which is equal to -0.0132 at baseline if $\varepsilon_{Sb} = 0.2$. In other words, in our case, each $1000 of tuition subsidy lowers fertility by about 0.013 children per person.

$e'(x_f)$, or the marginal effect of a change in fertility on social welfare, is much harder to evaluate convincingly, as we have documented in this paper: there are plausible reasons for it to be large and negative (environmental externalities), and plausible reasons for it to be large and positive (population ethics considerations). In our analysis, it takes a value of $-25.13U_1(x)$ in the environmental externality case and $38.02U_1(x)$ in the population ethics case.[16] Therefore, for a researcher wanting to evaluate the impact of population externalities on an optimal social policy in a second-best world in which it is hard to directly target fertility, our recommendation is to incorporate a term corresponding to $e'(x_f)\frac{dx_f}{dp}$ into their welfare derivative, using an estimated value of $\frac{dx_f}{dp}$ from the relevant literature, and a range of plausible values for $e'(x_f)$ – perhaps corresponding to our range of values of approximately $\{-25,40\} \times U_1(x)$.

If, as discussed above, the externality was also a function of other choices x_j, we would have to add additional terms to the welfare derivative. For example, if the externality was $e(x_f, x_c)$ where x_c represents inputs into child quality, then instead of simply adding $e'(x_f)\frac{dx_f}{dp}$ to the equation for $\frac{dW}{dp}$, we would have to add $\frac{\partial e(x_f,x_c)}{\partial x_f}\frac{dx_f}{dp} + \frac{\partial e(x_f,x_c)}{\partial x_c}\frac{dx_c}{dp}$. In the case of environmental externalities, this could occur if, for example, education led to an increase in the emissions of the parent or the child; in the case of imperfect parental altruism, education might generate a positive externality through x_c if education raises inputs into child quality. In each case, the conclusions of our welfare analysis would become more ambiguous, but we implicitly assume that $\frac{\partial e(x_f,x_c)}{\partial x_c}\frac{dx_c}{dp}$ is small relative to $\frac{\partial e(x_f,x_c)}{\partial x_f}\frac{dx_f}{dp}$ throughout, to focus on the fertility margin.

5. Conclusion

Our goal in this paper has been to demonstrate that public economists can incorporate population externalities in their analyses of public policy, and that the impact of such externalities can be highly significant. If fertility – and thus population – is not chosen optimally, and if policy measures to directly target fertility are costly and inefficient, then optimal policies in other areas could be strongly affected. We provide a demonstration for the policy of tuition subsidies for college: if such subsidies increase enrollment and graduation, and if that increase in education has a negative causal effect on fertility, then there could be an important population externality channel in our evaluation of the optimal subsidy.

We show that the optimal tuition subsidy is very sensitive to the population externalities we consider: optimal tuition subsidies should be much larger if we want to use education policy to help combat climate change, and optimal tuition subsidies should be much smaller if we want to use education policy to correct

[16] Our final welfare derivative in sections 2 and 3 was normalized into dollars rather than utils, and so we effectively measure $e'(x_f)$ as a constant multiplied by the $U_1(x)$ used for normalization.

imperfect parental altruism. We do not know which of these – if either – is the right answer, but we have shown that answering that question, and evaluating the sign and magnitude of population externalities, is very important.

The final section of the paper demonstrates how to generalize our approach to a broader set of policy questions: to any sufficient-statistic analysis of optimal policy, a new term can be incorporated into the welfare derivative that can be expressed as the product of the effect of the policy on fertility and the effect of that fertility change on welfare.

References

Acemoglu, Daron, and Jörn-Steffen Pischke. 2001. "Changes in the Wage Structure, Family Income, and Children's Education." European Economic Review 45 (4-6): 890–904.

Amin, Vikesh, and Jere R. Behrman. 2014. "Do More-Schooled Women Have Fewer Children and Delay Childbearing? Evidence from a Sample of US Twins." Journal of Population Economics 27 (1): 1–31.

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes. 2008. "Staying in the Classroom and Out of the Maternity Ward? The Effect of Compulsory Schooling Laws on Teenage Births." Economic Journal 118 (530): 1025–54.

Blanchet, Didier, and Olivia Ekert-Jaffé. 1994. "The Demographic Impact of Fertility Benefits: Evidence from a Micro-Model and from Macro-Data." In The Family, the Market and the State in Ageing Societies, edited by John Ermisch & Naohiro Ogawa, 79–104. Clarendon Press, Oxford.

Bruderer Enzler, Heidi, and Andreas Diekmann. 2015. "Environmental Impact and Pro-Environmental Behavior: Correlations to Income and Environmental Concern." Sociology Working Paper no. 9. ETH Zurich.

Cameron, Stephen J., and James J. Heckman. 2001. "The Dynamics of Educational Attainment for Black, Hispanic, and White Males." Journal of Political Economy 109 (3): 455–99.

Cannonier, Colin. 2014. "Does the Family and Medical Leave Act (FMLA) Increase Fertility Behavior?" Journal of Labor Research 35 (2): 105–32.

Card, David, and Thomas Lemieux. 2001. “Dropout and Enrollment Trends in the Postwar Period: What Went Wrong in the 1970s?” In Risky Behavior Among Youths: An Economic Analysis, edited by Jonathan Gruber, 439–82. University of Chicago Press.

Chetty, Raj. 2009. “Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods.” Annual Review of Economics 1: 451–87.

Climate Watch. 2019. “United States.” Https://www.climatewatchdata.org/countries/USA?calculation=PER_CAPITA. World Resources Institute.

Currie, Janet, and Hannes Schwandt. 2014. “Short- and Long-Term Effects of Unemployment on Fertility.” PNAS 111 (41): 14734–39.

Deming, David, and Susan Dynarski. 2009. “Into College, Out of Poverty? Policies to Increase the Postsecondary Attainment of the Poor.” Working Paper no. 15387. NBER, Cambridge, MA.

Dynarski, Susan. 2008. “Building the Stock of College-Educated Labor.” Journal of Human Resources 43 (3): 576–610.

EPA. 2017. “The Social Cost of Carbon: Estimating the Benefits of Reducing Greenhouse Gas Emissions.” Https://19january2017snapshot.epa.gov/climatechange/social-cost-carbon_.html. United States Environmental Protection Agency.

Epple, Dennis, Richard Romano, and Holger Sieg. 2006. “Admission, Tuition, and Financial Aid Policies in the Market for Higher Education.” Econometrica 74 (4): 885–928.

Fort, Margherita, Nicole Schneeweis, and Rudolf Winter-Ebmer. 2011. “More Schooling, More Children: Compulsory Schooling Reforms and Fertility in Europe.” Quaderni – Working Paper DSE No. 787. Università di Bologna Department of Economics.

———. 2016. “Is Education Always Reducing Fertility? Evidence from Compulsory Schooling Reforms.” Economic Journal 126 (595): 1823–55.

Gautier, Anne H. 2007. “The Impact of Family Policies on Fertility in Industrialized Countries: A Review of the Literature.” Population Research and Policy Review 26 (3): 323–46.

Government Offices of Sweden. 2020. “Sweden’s Carbon Tax.” Https://www.government.se/government-policy/taxes-and-tariffs/swedens-carbon-tax/.

Gruber, Jon, and Emmanuel Saez. 2002. “The Elasticity of Taxable Income: Evidence and Implications.” Journal of Public Economics 84 (1): 1–32.

Harford, Jon D. 1997. “Stock Pollution, Child-Bearing Externalities, and the Social Discount Rate.” Journal of Environmental Economics and Management 33 (1): 94–105.

———. 1998. “The Ultimate Externality.” American Economic Review 88 (1): 260–65.

Hendren, Nathaniel, and Ben Sprung-Keyser. 2020. “A Unified Welfare Analysis of Government Policies.” Quarterly Journal of Economics forthcoming.

Hyatt, Douglas E., and William J. Milne. 1991. “Can Public Policy Affect Fertility?” Canadian Public Policy 17 (1): 77–85.

James, Jonathan, and Sunčica Vujić. 2019. “From High School to the High Chair: Education and Fertility Timing.” Economics of Education Review 69: 1–24.

Kan, Kamhon, and Myoung-Jae Lee. 2018. “The Effects of Education on Fertility: Evidence from Taiwan.” Economic Inquiry 56 (1): 343–57.

Lavy, Victor, and Alexander Zablotsky. 2015. “Women’s Schooling and Fertility Under Low Female Labor Force Participation: Evidence from Mobility Restrictions in Israel.” Journal of Public Economics 124 (4): 105–21.

Lawson, Nicholas. 2017. “Liquidity Constraints, Fiscal Externalities, and Optimal Tuition Subsidies.” American Economic Journal: Economic Policy 9 (4): 313–43.

León, Alexis. 2004. “The Effect of Education on Fertility: Evidence from Compulsory Schooling Laws.”

Milligan, Kevin. 2005. “Subsidizing the Stork: New Evidence on Tax Incentives and Fertility.” Review of Economics and Statistics 87 (3): 539–55.

Monstad, Karin, Carol Propper, and Kjell G. Salvanes. 2008. “Education and Fertility: Evidence from a Natural Experiment.” Scandinavian Journal of Economics 110 (4): 827–52.

Moser, Stephanie, and Silke Kleinhückelkotten. 2018. "Good Intents, but Low Impacts: Diverging Importance of Motivational and Socioeconomic Determinants Explaining Pro-Environmental Behavior, Energy Use, and Carbon Footprint." Environment and Behavior 50 (6): 626–56.

OECD Family Database. 2019. "Sf2.3: Age of Mothers at Childbirth and Age-Specific Fertility." Https://www.oecd.org/els/soc/SF_2_3_Age_mothers_childbirth.pdf. OECD Directorate of Employment, Labour; Social Affairs.

Phipps, Shelley A. 2000. "Maternity and Parental Benefits in Canada: Are There Behavioural Implications?" Canadian Public Policy 26 (4): 415–36.

Ricke, Katharine, Laurent Drouet, Ken Caldeira, and Massimo Tavoni. 2018. "Country-Level Social Cost of Carbon." Nature Climate Change 8 (10): 895–900.

Robins, Philip K., and Paul Fronstin. 1996. "Welfare Benefits and Birth Decisions of Never-Married Women." Population Research and Policy Review 15 (1): 21–43.

Tropf, Felix C., and Jornt J. Mandemakers. 2017. "Is the Association Between Education and Fertility Postponement Causal? The Role of Family Background Factors." Demography 54 (1): 71–91.

Wei, Christina Chang, Lutz Berkner, Shirley He, Stephen Lew, Melissa Cominole, and Peter Siegel. 2009. "2007-08 National Postsecondary Student Aid Study (NPSAS:08): Student Financial Aid Estimates for 2007-08: First Look (NCES 2009–166)." Washington, DC: National Center for Education Statistics, Institute of Education Sciences, US Department of Education.

John Broome[1]

How Much Harm Does Each of Us Do?[2]

This paper estimates the amount of harm an average American does by her emissions of greenhouse gas, on the basis of recent very detailed statistical analysis being done by a group of economists. It concentrates on the particular harm of shortening people's lives. The economists' work presents the 'mortality cost' of emissions in terms of money. The actual quantity of life lost to climate change is embedded in their figures, but not in a transparent way. I have extracted a very rough and tentative estimate of it, by reverse engineering from their results. The estimate varies greatly according to how effectively the world responds to climate change. If the world's response is very weak, I estimate that an average American's emissions shorten lives by six or seven years in total. If the response is moderately strong, my figure is about half a year.

[1] University of Oxford, john.broome@philosophy.ox.ac.uk.

[2] My thanks for very useful comments, information and advice from Mark Budolfson, Tamma Carleton, Douglas MacLean, Tristram McPherson and John Nolt. Research for this paper was supported by ARC Discovery Grants DP140102468 and DP180100355. Financial support by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences) is gratefully acknowledged.

1. Sorts of harm and their quantity

Several moral philosophers have argued that the greenhouse-gas emissions of a single individual do no harm. I think they are mistaken, and I have opposed their arguments in a paper I called 'Against denialism'. Now I shall give some positive account of the quantity of harm that each of us does.

Many accounts already exist. First, there is a very large literature in economics on the 'social cost of carbon' (SCC), which is supposed to measure in terms of money the harm done by a tonne of carbon dioxide. Multiply the SCC by the total number of tonnes emitted by a person during her lifetime, and we get a money value for the total harm she does.

Estimates of the SCC vary greatly, but the average of all the estimates surveyed in a comprehensive meta analysis is $55.[3] This implies that a person who emits 1200 tonnes, which is typical for an American, causes $66,000 of harm. But a lot is left out of calculations of the SCC.[4] Anything whose value cannot be made commensurate with money is inevitably omitted or poorly taken into account. This includes the wellbeing of animals and whatever intrinsic value natural objects – such as natural species and individual trees – may have. It may also include human cultural goods, such as the culture of Arctic peoples and monuments that may be lost to the sea.

Climate change kills very many people. It kills them directly in droughts, floods, and storms. It kills them less directly through increasing the range of tropical diseases and by impoverishing the people who struggle to live in less hospitable parts of the world. Estimates of the SCC in principle take into account the harm of killing, but in practice they generally do so badly. The value they assign to the loss of a life is generally based on people's willingness to pay for extending their lives, and it does not properly recognize the different value of money to different people. Poor people are willing to pay less that rich people to reduce their risk, but this is not because their lives are less valuable. It is because they have other, more pressing uses for their money. Economists often ignore this simple point.[5] Also, they generally discount later lives compared with earlier ones.

Some evidence will emerge in section 2 that killing is a major part of the harm that climate change will do. Since existing estimates of the SCC take it into account badly, it is important to pay special, separate attention to this harm. The SCC conveys some information about the harm done by climate change, but we separately need information about killing. This paper will focus on estimating the quantity of killing we do through climate change.

[3] Wang Pei et al 'Estimates of the social cost of carbon'.

[4] See the full discussion in Fleurbaey et al, 'The social cost of carbon.'

[5] The IPCC asks them not to ignore it. See IPCC, *Climate Change 2014,* Summary for Policymakers, p. 5.

There are precedents for this estimation too. One figure for the amount of this harm is already frequently quoted in moral philosophy. It originates from calculations that John Nolt published in 2011.[6] Nolt started by working out carefully that an average American is responsible for about the fraction 5 x 10–10 of the climate change caused by greenhouse gas emissions up to 2040. Next he calculated the number of people who will live during the next millennium as 100 billion. Then he says:

> If over the next millennium as few as four billion people (about 4% [of the number who will live during that period]) are harmed (that is, suffer and/or die) as a result of current and near term global emissions, then the average American causes through his/her greenhouse gas emissions the serious suffering and/or deaths of two future people.[7]

Nolt did not try to justify the figure of 4% for the proportion of people who will be harmed. He was not aiming to estimate the amount of harm so much as to illustrate what it might be. Nevertheless 'the serious suffering and/or deaths of two future people' is frequently quoted as his estimate – recently in the *New York Times*.[8] We shall see that, as an estimate, it is far too big.

At about the same time, I published an estimate based on figures from the World Health Organization.[9] My estimate was that a typical westerner takes away more than six months of human life altogether.[10]

2. New data and estimates

Those figures are now very much out of date. Much better ones are becoming available. A major report, 'Valuing the Global Mortality Consequences of Climate Change' (VGMC), derives conclusions on the basis of extremely extensive and detailed data about the effect of weather on death rates at a very local level. The authors divided the land surface of the Earth into 24,378 areas and assembled data on a 38% sample of them. By means of sophisticated statistical analysis, they have derived from their data authoritative estimates of the global effects of climate change on mortality. I shall base my conclusions on these estimates.

To put it very crudely, VGMC regresses death rates on temperature. This means

[6] Nolt, 'How harmful are the average American's greenhouse gas emissions'.

[7] p. 9.

[8] Newman, 'If seeing the world helps ruin it, should we stay home?' Thanks to Douglas MacLean for this reference.

[9] WHO, *Global Health Risks*.

[10] *Climate Matters*, p. 74.

it takes account of all causes of death – all the various means by which the warming of the planet kills people. It includes deaths in heat waves, deaths resulting from the spread of tropical diseases, and so on. It is not limited to particular causes of death, as are earlier studies from the World Health Organization.[11]

VGMC also takes account of the ages of people who die, so it can calculate the number of life-years lost as well as crude death rates. These are much more informative. Many people die in heat waves, and this is one of the significant ways in which climate change kills people. But many of those people are elderly[12] and many are already suffering from chronic diseases. A heat wave may shorten their lives by only a few years, months or days. Climate change also increases the prevalence of diarrhoeal diseases; this is another significant means by which it kills people. It is mainly children who die from these diseases,[13] and they lose many years of life. It would be misleading to count a child's death and an elderly person's death the same, and VGMC does not do so. All in all, VGMC data is very valuable.

Still, estimates of harm from climate change can never be certain. The science of climate change is very uncertain, and the spread of possibilities is very wide. For example, it is possible that climate change will lead to a catastrophe for humanity, and even to our extinction. It may even be that in responding to climate change we should care more about this unlikely possibility of catastrophe than about what is likely to happen. Quite generally, an unlikely but very bad event may be more important for our planning than what is likely to happen. That is why ships ought to carry lifeboats. A ship is unlikely to sink, so its lifeboats are unlikely to be used. But if it does sink the consequences of having no lifeboats will be so dire that they make the expense of carrying lifeboats worthwhile. The economist Martin Weitzman argues that our response to climate change should be like our response to the possibility of a ship's sinking: directed towards the unlikely but very bad consequences of catastrophe.[14]

I have no way to estimate the harm that a person's emissions will do if the results of climate change are catastrophic. I am therefore forced to limit my estimates to the amount of harm that is likely.

Even the likely harm done by a person's emissions is very contingent; it depends on the emissions of other people. This is because the relation between temperature

[11] For example, WHO, *Global Health Risks* and *Quantitative Risk Assessment*.

[12] WHO, *Quantitative Risk Assessment*, p. 17.

[13] WHO, *Quantitative Risk Assessment*, p. 37.

[14] Weitzman, 'On modeling and interpreting the economics of catastrophic climate change'. See also Wagner and Weitzman, *Climate Shock*. My paper 'The most important thing about climate change' explains that Weitzman's argument is insufficient for his conclusion.

and mortality is very non-linear. Its graph is U-shaped. Both low temperatures and high temperatures cause an increased number of deaths. As the temperature increases starting from a low level, the death rate decreases until it reaches a minimum at around 20C. Then it starts to increase at an accelerating rate. Consequently, an increase in temperature when the temperature is very high causes much more harm than the same increase would do were the initial temperature lower. An emission of greenhouse gas causes much more harm if other emissions are high than if they are low.

This contingency is handled in climate-change science by means of 'scenarios'. Each scenario describes a particular possible future development of emissions together with the growth of the world's population and economy. So when I refer to the harm that is likely to result from a person's emissions, I mean the harm that is likely given a particular scenario. The VGMC study works with two scenarios known as 'RCP 4.5' and 'RCP 8.5'.[15] Perforce, I copy it in this respect. RCP 4.5 is a moderate scenario in which emissions of greenhouse gases begin to decline around the middle of this century. Nevertheless, the temperature under RCP 4.5 is likely to reach 2.4C above pre-industrial levels, which is well above the target set in the Paris Agreement negotiated in 2015 by the United Nations Framework on Climate Change. So this is by no means an optimistic scenario. RCP 8.5 is intended to be a baseline that might be considered 'business as usual'. It should be treated as a basis for comparison rather than a prediction of what will happen. It assumes high growth of population with slow economic growth, and limited technical progress. In RCP 8.5, emissions increase through the century, and the temperature is expected to reach almost 5C above pre-industrial levels. This might fairly be counted as catastrophic. RCP 8.5 is a very pessimistic scenario.

VGMC calculates what it calls the 'mortality-related' harm that will result from emitting one tonne of carbon dioxide in 2020. By means I shall explain, it expresses the result in terms of dollars. To cut a long story short, its conclusion is that the dollar value of the harm is $18.9 under RCP 4.5 and $98.9 under RCP 8.5. (VGMC p. 46.) These are the figures I shall work with. They assume a 2.5% discount rate on commodities, and a 'globally uniform valuation of mortality risk'. I shall explain these two assumptions in sections 4 and 3 respectively.

Compare these figures with $50, which is a typical estimate for the SCC as a whole. The comparison supports the assertion I made in section 1, that mortality-related harm is at least a major part of the harm that climate change will do.

[15] See Wayne, 'The beginner's guide'.

3. Lives for money

VGMC presents its conclusions in terms of money values. But many philosophers including me are dubious about translating the value of lives into money. We would prefer to see the result in terms of quantities of life itself. This raw information is embedded in the VGMC calculations, but not in a transparent way. The authors of the report are in a position to extract and present it, but only by means of a substantial amount of computation. I believe they will do so in due course.

In the meantime this volume goes to press. In order to give readers some rough idea of the quantity of life we take away through our emissions, I have extracted estimates of this quantity from the figures presented in the existing report, using the best means I have available. These means are frankly very crude. The outcome will be very approximate, but it is the best I can do. The authors of the report bear no responsibility for my figures, and mine will be totally superseded by theirs when they are published.

I first adjust the figures by subtracting adaptation costs from them. As temperatures increase, people adapt to them. Their bodies acclimatize and they take steps to avoid the heat. VGMC uses sophisticated methods to account for adaptation in its estimates of the number of people killed by climate change. It also recognizes that adaptation often costs money, and it includes this cost in its figures for mortality-related harm. I want to estimate the actual amount of killing that climate change does, so I need to subtract the adaptation costs. VGMC states that on average 14% of mortality-related costs are adaptation costs (VGMC, p. 5). I have to use this average figure because I cannot find figures in VGMC related to the particular costs I am working with. I therefore reduce those costs by 14% and get $16.3 for RCP 4.5 and $83.1 for RCP 8.5. These are now the dollar values of life actually lost.

Next I work back from these dollar values to calculate the actual quantities of life that they represent. The dollar values are based on the monetary value of life that is standardly used in cost-benefit analysis in the US. This is $10.95 million for a life (VGMC p. 121). VGMC converts it to a value for a year of life by using the life expectancy of a median-aged American (VGMC p. 120). The text does not state what this life-expectancy is, so I have to recover it. The median age of Americans in 2018 was 38.2.[16] Life tables for 2016 show life expectancy at 38.2 as 40.23 for men and 44.20 for women.[17] I shall assume an average life expectancy of 42.2. The result is that a life-year is valued at $259,000.

In principle, VGMC values lives or life-years on the basis of what people are willing to pay for them, or more exactly what they are willing to pay to improve their

[16] https://www.statista.com/statistics/241494/median-age-of-the-us-population/

[17] https://www.ssa.gov/oact/STATS/table4c6.html.

chances of living longer. It assumes that what people are willing to pay is proportional to their income (VGMC, p. 39). In practice, willingness to pay is averaged across the population. The value of $259,000 is the average across the US population. The VGMC figures for 'globally uniform valuation of mortality risk' are based on average willingness-to-pay across the whole world population, under the assumption that willingness to pay is proportional to income. This is what 'globally uniform' means. The consequence is that the figure of $259,000 needs to be reduced by the ratio of global average income to American average income. From World Bank data[18] in 2018 I find this ratio to be .287. This makes the value of a life year $74,300.

This value allows us to take the above-quoted dollar values of killing caused by a tonne of carbon dioxide and convert them back into numbers of life years. The result is .000219 life years in RCP 4.5 and .00112 in RCP 8.5.

Those are rates per tonne of carbon dioxide. Next we have to multiply these quantities by the number of tonnes of carbon dioxide emitted by a person during her lifetime. There is a further complication here. These quantities I have derived from VGMC measure the harm done by a tonne of carbon dioxide that is emitted in 2020. Later emissions are done at a time when the global temperature is higher. Because of the non-linear relationship between temperature and deaths, later emissions therefore do more harm. Earlier emissions do less harm. People who are young now will do more harm by their emissions than older people who emit the same in total. Because I do not aim at precision, I shall ignore this complication. I shall consider a person who emits 1200 tonnes during her lifetime, which is about average for an American. Multiplying the rates of death per tonne by this amount, we might conclude that this person's lifetime emissions cause the loss of .263 life years in RCP 4.5 and 1.34 life years in RCP 8.5.

4. The consequences of discounting

Sadly, this is still not correct. These figures are a serious underestimate because they incorporate some discounting of life-years. So I turn to the difficult issue of discounting.

The VGMC figures assume a discount rate of 2.5% on commodities. Strictly, this is the discount rate on money values after cancelling out inflation. This means it is the discount rate on the bundle of commodities that are used as the basis for measuring inflation. It is correct to discount future commodities – which means giving less value to future commodities than to present ones – because both scena-

[18] https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD.

rios RCP 4.5 and RCP 8.5 assume that economic growth will continue. That is to say, they assume people will be becoming progressively richer. Therefore, the value of commodities to them at the margin (the value of adding to their stock of commodities) will progressively diminish. This is the consequences of the diminishing marginal value of income, which has been recognized in economics at least since the time of Alfred Marshall. The more commodities you are already consuming, the less you will benefit from consuming more commodities.

VGMC does not discount life-years at 2.5%, because it gives a progressively increasing value to future life-years. I have not mentioned this before. The reason it does so is that it takes the value of life-years to be proportional to income, and income increases with economic growth. The average rate of economic growth through this century is 2% per year in RCP 4.5 and 1.35% in RCP 8.5.[19] The value of life-years is increased at these rates. So in effect VGMC discounts life-years at a rate equal to the difference between these rates and 2.5%. The upshot is that the figures for life-years I have been working with incorporate discount rates of 0.5% for RCP 4.5 and 1.15% for RCP 8.5.

The reason I gave for discounting commodities does not apply to life-years. It is most implausible that life-years lived later in history are really less valuable than ones that are lived earlier. Discounting life-years is an instance of what is called 'pure discounting'. Pure discounting has received some support from economic theorists – notably Kenneth Arrow – but not much.[20] It does not sit well with the globally uniform valuation embedded in the VGMC figures I have been using. If, at each date, everyone's life-years are given the same value, why should life-years at one date be given a different value from those at another date?

Moreover, discounting as VGMC does it has a peculiar consequence. The loss of a person's life-year is discounted according to the date of the person's death rather than the date when the life-year would have been lived. Suppose a 20-year-old dies now and loses 60 years of life. The loss of the years she would have lived between the ages of 60 and 80 gets full value in the calculation. But if someone else born at the same time as this 20-year-old lives to 60 and then dies, losing 20 years of life, those 20 years are discounted relative to the 20-year-old's, even though they would have been lived at exactly the same time.

A life-year plausibly has a constant value, whoever lives it and whenever it is lived. Because of this, it provides a plausible basis for inter-personal comparisons of the value of commodities, as the IPCC explains.[21] It also provides a plausible basis

[19] I derive these figures from Figure 12 of Wayne 'The Beginner's Guide'.

[20] Arrow, 'Discounting, morality and gaming'. See the discussion in my 'The wellbeing of future generations'.

[21] IPCC, *Climate Change 2014,* box 3, p. 226.

for the inter-temporal value of commodities that appear in the discount rate. Plausibly, commodities should be discounted at whatever rate implies a constant value for a year of life. If we maintain VGMC's assumption that the value of a life-year is proportional to income, this means discounting commodities at the rate of growth of income. That is: 2% in RCP 4.5 and 1.35% in RCP 8.5.

I urge the authors to provide figures for years of life lost corresponding to these discount rates. At least, they should not treat the discount rate in the way they do, as exogenously given independently of the growth rate. The correct discount rate is a function of the growth rate, as the famous Ramsey equation shows.[22]

The estimates obtained at the end of section 3 incorporate discount rates on life-years of 0.5% in RCP 4.5 and 1.15% in RCP 8.5. In the absence of undiscounted figures in VGMC, I need to cancel out the discounting as best I can. How badly do the discounted estimates underestimate the true quantity? This depends on how the killing caused by an emission of carbon dioxide is distributed over time. This information is implicit in the work of VGMC, but I cannot extract it from the paper. I can make only guesses.

When a tonne of carbon dioxide is emitted, it causes the atmosphere's temperature to rise soon afterwards, and that raises the death rate. The tonne begins immediately to be absorbed by the land and oceans, so its effect on the death rate will begin to fall too. About half the tonne will fall out of the atmosphere within 50 years. However, perhaps 20% of it will persist for hundreds and even thousands of years.[23] Furthermore, its effect on temperature will lag behind the quantity of carbon dioxide itself. I am not in a position to judge the extent of the lag; doing so would require running a model of the atmosphere.[24] But it is plain that, were the killing to decline only with temperature, it would continue for a very long time and its total would be very large.

But actually the killing will be progressively reduced by people's adaptation to the heat. The VGMC data contains some information about adaptation, but I have not been able to use it at this point in the calculation. In any case, it could tell us very little about the development of humanity several centuries from now. Human life will be so different in three hundred years that it is hard to know even how adaptation could be identified by then. I have to fall back on little more than guesswork. Bearing in mind that the quantity of an emitted tonne of carbon dioxide will be reduced by about a half in half a century, I shall assume that its effect on the death rate will be reduced to half within a century. I assume that the effect on the death

[22] IPCC, *Climate Change 2014*, p. 229.

[23] See the graphs in box 6.1 on p. 473 of IPCC, *Climate Change 2013: the Physical Science Basis*.

[24] IPCC, *Climate Change 2013,* pp. 1102B5.

rate will be very small after three centuries. (Three centuries is the horizon set on the calculations in VGMC.)

For RCP 4.5, the estimate I obtained at the end of section 3 for the life-years taken away by a person who emits 1200 tonnes is .263. The discount rate is 0.5%. This amounts to a discount of 40% over 100 years, 63% over 200 years, and 78% over 300 years. This suggests to me that discounting at this rate is not likely to underestimate the total of harm by more than 50%. I guess therefore that in RCP 4.5 the amount of killing done by a person who emits 1200 tonnes is in the region of half a life-year.

For RCP 8.5, the estimate I obtained at the end of section 3 is 1.34 life-years. The discount rate is 1.15%, which amounts to 68% over 100 years, 89% over 200 years and 96% over 300 years. This suggests an underestimate of perhaps 75% or 80% in the amount of killing done. I guess that in RCP 8.5 the amount of killing done by a person who emits 1200 tonnes is in the region of 6 or 7 life-years.

5. Conclusion and why it matters

My attempt in section 4 to cancel out discounting from the figures is the most speculative part of my calculation. I was forced to speculate about the adaptive success of human beings centuries in the future. Since the VGMC estimates for RCP 4.5 imply only the small discount rate on life of 0.5%, they are less vulnerable to a mistake about this. The higher discount rate of 1.15% implied in the estimates for RCP 8.5 makes a much greater difference. Remember that in any case RCP 8.5 does not represent a prediction so much as a worst-case baseline. RCP 4.5 is more like a prediction, and I put much more trust in the RCP 4.5 figures.

There is anyway a great deal of uncertainty in any quantitative predictions involving climate change. I have tried to work out only the harm that is likely to arise from emissions; much greater harm is possible. Furthermore, remember I am only trying to produce interim results, while I wait in hope that the authors of VGMC will produce much more accurate ones in due course. With all these caveats, my best estimate of the amount of life you are likely to take away by emitting 1200 tonnes of carbon dioxide is half a year.

Why does it matter? It helps to position the harm we do through climate change on the scale of all the good and bad things we do. It is important to recognize that the harm an individual does by her emissions – and correspondingly the good she can do by reducing her emissions – though definitely significant, is not large in comparison to other means of doing good.

Some ways of reducing emissions, such as eating less meat and turning down the air conditioning, are easy and cheap. Others, such as insulating your house, are expensive. One of the cheaper ways is to offset your emissions. You can offset by planting trees or by paying for projects that reduce emissions elsewhere. The cost of offsetting is in the region of $10 per tonne. According to my figures, if you were to spend $12,000 on offsetting your lifetime emissions of 1,200 tonnes, you would save perhaps half a life-year in RCP 4.5 and 6 or 7 life years in RCP 8.5. By contrast, the organization GiveWell lists on its website charities that, on its calculations, can save a person's whole life for a donation of $2,000 or $3,000.[25] Among them are charities that fight malaria. These are plainly more effective ways of using money to do good.

Why, then, should you reduce your emissions? Mainly because justice requires it. You emit greenhouse gas to benefit yourself, but in doing so you harm other people. It is an elementary principle of common-sense justice that, with certain exceptions such as self-defence, you should not harm other people for your own benefit. On this point I agree with Nolt.[26]

It is also true that climate change is in aggregate doing immense harm in the world. Although reducing emissions is not the most effective way of doing good, it is well worth the cost. For you as an individual, this is not unqualifiedly so because you have better ways of using your money. If you use your money in the best ways, starting with the best means of doing good and working down to less good means, you will run out of money long before you get to reducing your emissions much. But governments are different because they control vastly greater resources. They have coercive power over their people's behaviour, by means of taxes and regulations. It is true for a government as it is for an individual, that it should first direct resources in more effective ways such as fighting malaria. But when all that is done, a government should still direct a vast amount of further resources towards reducing emissions of greenhouse gas.

An appropriate means of doing good for an individual is therefore political action aimed at getting governments to reduce emissions. This is a further reason for reducing your own emissions. Doing so is a sort of political action. It shows that you care. It may induce others to follow you and to vote for reducing emissions.

References

Arrow, Kenneth, 'Discounting, morality, and gaming', in Discounting and Intergenerational Equity, edited by P. R. Portney and J. P. Weyant, Resources for the Future, 1999, pp, 13–21.

[25] https://www.givewell.org/how-we-work/our-criteria/cost-effectiveness/cost-effectiveness-models

[26] 'How harmful are the average American's greenhouse gas emissions?', p. 7. See my *Climate Matters*, chapter 4.

Broome, John, ‘Against denialism’, The Monist, 102 (2019), pp. 110–29.

Broome, John, Climate Matters: Ethics in a Warming World, Norton, 2012.

Broome, John, ‘The most important thing about climate change’, in Public Policy: Why Ethics Matters, edited by Jonathan Boston, Andrew Bradstock and David Eng, ANU E Press, 2010, pp. 101–116.

Broome, John, ‘The well-being of future generations’, in The Oxford Handbook of Well-Being and Public Policy, edited by Matthew Adler and Marc Fleurbaey, Oxford University Press, 2016, pp. 901–28.

Fleurbaey, Marc, Maddalena Ferranna, Mark Budolfson, Francis Dennig, Kian Mintz-Woo, Robert Socolow, Dean Spears and Stéphane Zuber, ‘The social cost of carbon: valuing inequality, risk and population for climate policy’, The Monist, 102 (2019), pp. 84–109.

Intergovernmental Panel on Climate Change, Climate Change 2013: the Physical Science Basis, Cambridge University Press, 2013.

Intergovernmental Panel on Climate Change, Climate Change 2014: Mitigation of Climate Change, Cambridge University Press, 2014.

Newman, Andy, ‘If seeing the world helps ruin it, should we stay home?’, New York Times, 3 June 2019.

Nolt, John, ‘How harmful are the average American’s greenhouse gas emissions?’, Ethics, Policy & Environment, 14 (2011), pp. 3–10.

Carleton, Tamma, Michael Delgado, Michael Greenstone, Trevor Houser, Solomon Hsiang, Andrew Hultgren, Amir Jina, Robert Kopp, Kelly McCusker, Ishan Nath, James Rising, Ashwin Rode, Hee Kwon Seo, Justin Simcock, Arvid Viaene, Jiacan Yuan, and Alice Zhang, ‘Valuing the Global Mortality Consequences of Climate Change Accounting for Adaptation Costs and Benefits’, 2019.

Wagner, Gernot, and Martin Weitzman, Climate Shock: The Economic Consequences of a Hotter Planet, Princeton University Press, 2015.

Wang, Pei, Xiangzheng Deng, Huimin Zhou and Shangkun Yu, ‘Estimates of the social cost of carbon: A review based on meta-analysis’, Journal of Cleaner Production, 209 (2019), pp. 1494–1507.

Wayne, Graham, ‘The Beginner’s Guide to Representative Concentration Pathways’, https://skepticalscience.com/docs/RCP_Guide.pdf.

Weitzman, Martin L., ‘On modeling and interpreting the economics of catastrophic climate change’, Review of Economics and Statistics, 91 (2009), pp. 1–19.

World Health Organization, Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks, 2009.

World Health Organization, Quantitative Risk Assessment of the Effects of Climate Change on Selective Causes of Death, 2030s and 2050s, 2014.

Tim Campbell[1]

Offsetting, Denialism, and Risk[2]

In *Climate Matters* John Broome defends two claims. First, if you live a "normal life" in a rich country, you will probably cause significant harm by your emissions of greenhouse gas (GHG), violating a moral duty of harm-avoidance. Second, you can satisfy this duty by offsetting your emissions. Some would deny Broome's first claim on the grounds that an individual's emissions of GHG do no harm. Broome calls this position "Individual Denialism" (ID) and in a recent paper he attempts to refute it. I explain how, if Broome's refutation of ID were successful, it would undermine his claim that you can satisfy your duty of harm avoidance by offsetting. I suggest an alternative defence of the claim that you can satisfy your individual duty to reduce your carbon footprint by offsetting. This alternative defence assumes that your duty to reduce your carbon footprint derives from a duty of risk-avoidance.

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

[2] For very helpful discussion of the issues in this paper, I would like to thank Vuko Andric, Gustaf Arrhenius, Dan Baras, Katharina Berndt Rasmussen, Paul Bowman, John Broome, Gunnar Björnsson, Krister Bykvist, Hilary Greaves, Anders Herlitz, Karim Jebari, Julia Mosquera, Joe Roussos, Orri Stefánsson, and participants at the 2020 Stockholm Workshop in Philosophy. I am especially grateful to Håkan Salwén for extremely helpful written comments on an earlier draft of this paper. Financial support from Riksbankens Jubileumsfond (grant M17-0372:1) is gratefully acknowledged.

1. Introduction

Anthropogenic climate change causes harm. As global greenhouse gas (GHG) levels increase, global temperatures increase, and people in the developing world (and elsewhere) die from heat exhaustion, tropical disease, and a host of other causes. It seems that states have a moral duty to reduce the harm of climate change by reducing their emissions of GHG.[3]

But what does this mean for you? Do you have a moral duty to reduce your personal carbon footprint? John Broome (2012) claims that if you live a "normal life" in a rich country, you will probably do significant harm by your emissions of GHG. Specifically, he claims, your lifetime emissions will probably wipe out (roughly) more than six months of healthy life, and that you would thereby violate a moral duty to avoid inflicting harm.[4]

However, Broome also claims that if you offset your GHG emissions you do *no harm* by emissions, and hence, satisfy your individual duty of harm-avoidance (other things being equal). Offsetting your emissions means you ensure that for every amount of GHG you cause to enter the atmosphere, you cause that much GHG to be subtracted from the atmosphere (Broome 2012, p. 85). In theory, you could achieve this by funding one or more of several ongoing emissions reduction projects, such as conserving or planting forests, replacing energy-inefficient stoves in the developing world with cleaner ones, or building and maintaining wind farms. If the amount of GHG that you remove from, or prevent from entering, the atmosphere by funding some such project is at least as much as the amount you emit, then you offset your emissions; your carbon footprint is neutral.

However, some would deny that your moral duty of harm-avoidance requires that you offset, or even reduce, your GHG emissions. They claim that an individual's emissions of GHG do *no harm*.[5] Broome calls this claim Individual Denialism, and in a recent paper he attempts to show that it is almost certainly false.[6] In this paper, I

[3] Reducing global carbon emissions is one important goal. Another is to help the developing world adapt to environmental changes that climate change causes.

[4] This assumes you would emit 800 tonnes of carbon dioxide (Co2) over your lifetime, which is an estimate for the average person born in a developed country in 1960. The estimate of harm from this amount of Co2 emitted is derived from the social cost of carbon. The moral duty that Broome thinks you would violate if you were to emit this much Co2 does not derive from your status as a citizen of any state or as a member of any collective; it is a non-derivative individual moral duty.

[5] For arguments that an individual's emissions do no harm, see e.g. Cripps (2013, pp. 119–24) and Maltais, Aaron 2013. "Radically Non-Ideal Climate Politics and the Obligation to at Least Vote Green," Environmental Values 22: 589–608. The view that an individual's emissions of GHG do no harm is sometimes attributed to Walter Sinnott-Armstrong (2005). Although Sinnott-Armstrong never explicitly endorses this view, he does claim that certain activities, such as going for a joyride in a gas-guzzling SUV, do no harm and that there is no moral requirement to refrain from engaging in them.

[6] "Against Denialism" (2018).

explain how Broome's case against Individual Denialism undermines his claim that you can satisfy your duty of harm avoidance by offsetting your emissions. I then suggest an alternative defence of the claim that by offsetting your emissions you can satisfy your individual duty to reduce your carbon footprint. This alternative defence, which is consistent with Broome's argument against Individual Denialism, assumes that your individual duty to reduce your emissions derives from a duty of risk avoidance rather than from a duty of harm avoidance.

Section 2 of the paper discusses Broome's defence of offsetting in *Climate Matters* as well as his argument against Individual Denialism in a recent paper, and explains how his argument against Individual Denialism undermines his claim that you can satisfy your duty of harm-avoidance by offsetting. Section 3 suggests an alternative defence of the claim that you can satisfy your individual duty to reduce your carbon footprint by offsetting and considers some possible worries that a more developed version of this defence must address. Section 4 concludes.

2. Broome on Offsetting and Individual Denialism

2.1 Broome's defence of offsetting

In *Climate Matters* Broome claims that offsetting your emissions is sufficient to satisfy your duty to avoid causing harm by emissions. He acknowledges some important concerns about offsetting. For example, he admits that if offsetting were a widely accepted practice, this might cause countries to delay cutting their emissions, thus slowing progress on climate change mitigation (Broome 2012, p. 94). But Broome thinks that this concern is primarily about what governments should do, and his defence of offsetting is presented in the context of discussing what an ordinary individual's duties are.

In this context, one important concern about offsetting that Broome considers is based on the claim that offsetting cannot undo the harm that is done by whatever greenhouse gasses one causes to enter the atmosphere. This worry has been voiced by Greenpeace, an organization to which Broome attributes the following statement:

> The truth is, once you've put a tonne of CO2 into the atmosphere, there's nothing offsetting can do to stop it changing our climate.[7]

[7] The 2007 statement is quoted on page 89 of *Climate Matters* and is attributed to Charlie Kronick of Greenpeace.

The concern underlying Greenpeace's statement seems to be that offsetting one's emissions would not prevent harm done by whatever GHG one emits but would at best avert further harm from emissions. According to Greenpeace, you cannot avoid doing harm by emissions simply by offsetting.

Broome dismisses Greenpeace's statement as "disingenuous", and claims

> ... As far as the climate is concerned, emitting a tonne of carbon dioxide and offsetting it is exactly as good as not emitting it in the first place, providing that the offset is genuine.[8]

He also claims

> Offsetting does not remove the very molecules that you emit, but the climate does not care which particular molecules are warming it. If you successfully offset all your emissions, you do no harm by emissions. You therefore do no injustice by them.[9]

Here Broome relies on an unstated assumption about the relationship between GHG emissions and harm, namely that an individual inflicts harm by GHG emissions only if that individual increases *cumulative GHG emissions*.

To illustrate the assumption, consider an analogy. Suppose that some people are trapped in a large tank that is slowly filling with water. They are unable to escape, even with outside help. As the water level increases, they struggle to keep their heads above water. The more the water level rises, the more they struggle. If the water level becomes high enough, some of them will drown. Now suppose you come along and dump exactly one litre of water into the tank but then immediately remove exactly one litre. Overall, the water level is exactly what it would have been if you had done nothing. Even if the water level continues to rise to the point at which people drown, it may seem that your behaviour did not contribute to this harm, at least when we compare your behaviour to a baseline in which you have no interaction whatsoever with the people in the tank.[10] In *Climate Matters*, Broome seems to think of the relationship between an individual's emissions and the harm of climate change along similar lines. The accumulation of greenhouse gas in the atmosphere is like the rising water level in the example just described, and the victims of accumulating

[8] Broome (2012), p. 89.

[9] Broome (2012), p. 85.

[10] This assumption may be questioned, however. It seems *possible* for your behaviour of adding and removing water to cause harm. Perhaps one should say that it is *very unlikely* that this behaviour would cause harm.

GHG are like the victims trapped in the water tank; they suffer more harm (e.g. they struggle harder to obtain the food and water they need), and are subjected to greater risk of further harm (e.g. greater risk of death), as cumulative emissions increase.

We shall return to the topic of offsetting below. We must first look at a question that Broome addresses in his more recent work, namely whether an individual's GHG emissions do harm at all.

2.2 Broome on Individual Denialism

Whether you can satisfy your duty of harm avoidance by offsetting your emissions is an important question only if an individual's emissions do harm. But not everyone is convinced that an individual's emissions do harm. Sinnott Armstrong (2005) claims that the GHG emissions from your driving a gas guzzler for fun on some occasion would not contribute to the harm of global warming. He compares the addition of GHG from driving to pouring a quart of water into "a river that is going to flood downstream because of torrential rains" (2005, p. 298). He assumes that just as the extra quart of water would make no difference to the victims of the flood, the extra GHG from driving would make no difference to the victims of global warming. Maltais (2013) and Cripps (2013) seem to hold a similar view regarding an individual's lifetime emissions of GHG.

Broome addresses these arguments in his recent paper "Against Denialism" (2018). He refers to the claim that an individual's emissions of GHG do no harm as Individual Denialism, and the paper is devoted to showing that this claim is almost certainly false.

According to Broome, a central problem with arguments for Individual Denialism is that they ignore "the significance of the atmosphere's extreme instability" (Broome 2018, p. 110). Broome thinks that the best available meterological science supports the claim that the atmosphere is a "chaotic system", meaning that even very small changes in the state of the atmosphere at a given time, can and frequently do cause (or alter the course of) very large meteorological events, such as tropical storms, at other global locations several weeks (and months, years, decades, etc.) later. He cites a famous lecture by the mathematician and meteorologist Edward Lorenz in which Lorenz asks whether the flap of a butterfly's wings in Brazil can set off a tornado in Texas and concludes that "it might". Broome writes

> It remains an unresolved question in meteorology whether a disturbance as small as a butterfly-flap can really escalate to a global scale. If it cannot, the reason is that the disturbance of a butterfly-flap is on such a small scale that the viscosity of the air may damp out its effect. For the atmosphere, the scale on which viscosity is significant is less than a centimetre. (Broome, p. 112)

Because the scale on which air viscosity is significant is less than a centimetre, Broome seems convinced that the doubts about the butterfly effect do not arise in the case of most emissions of GHG. Emissions from acts such as driving a gas guzzler will dissipate an amount of energy into the atmosphere that is many times greater than that of a butterfly-flap, and over the next century will cause "more than a trillion joules of energy from the sun to be absorbed by the earth", some of which "will warm the atmosphere and continue to stir it up" (2018, p. 112).

How is this relevant to assessing the impact of an individual's emissions? Broome answers this question by focusing on the example of driving a gas guzzler. He asks readers to imagine that going for a joy ride in a gas guzzling SUV would emit 25 kilos of carbon dioxide. He claims

> Given the atmosphere's instability, we should expect global weather in a few decades' time to be completely different if you go joyguzzling … from what it would have been had you stayed home. … Increasing emissions does not cause continuous changes punctuated by occasional discrete events such as a typhoon or a child's death from cholera. … Instead it will cause typhoons to form at quite different times and places, and it will lead to a completely different distribution of cholera outbreaks. Your … drive will cause a completely different group of people to be exposed to cholera and other risks of death. Some who would have died will survive because of your drive, and others who would have survived will die. … There is literally zero probability that emitting 25 kilos will do no harm and no good. (Broome 2018: 112—113)

Broome is not claiming that the global climate is unpredictable; the climate consists of "long-run averages" of weather and is "much more stable and predictable" than weather. Rather, Broome is claiming that an individual's emissions of GHG that cause sufficiently large atmospheric disturbances will change weather patterns (e.g. local temperatures, floods, storms, etc.) over a long time horizon, ultimately leading to a distribution of harm that would have been different if the individual's behaviour had been different.

Apparently, Broome thinks that if the atmosphere is a chaotic system then Individual Denialism is almost certainly false. Although Broome's discussion in "Against Denialism" focuses mainly on refuting specific "denialist" arguments, it seems one can extrapolate a master argument against Individual Denialism from claims that he makes. In particular, Broome seems committed to what I will call *The Chaos Argument*:

(1) The atmosphere is a chaotic system.

(2) If the atmosphere is a chaotic system, then creating sufficiently large atmospheric disturbances almost certainly does harm (there is "literally zero probability" that it does no harm).[11]

(3) Your emissions of GHG (e.g. from driving a gas guzzler) create sufficiently large atmospheric disturbances.

Therefore,

(4) Your emissions of GHG almost certainly do harm.

Therefore,

(5) It is almost certainly false that your emissions do no harm.

Therefore,

(6) Individual Denialism (which entails that your emissions do no harm) is almost certainly false.

A crucial assumption underlying Individual Denialism is that your emissions do not make a big enough positive difference to cumulative GHG levels to inflict harm. (Think of Sinnott-Armstrong's example of pouring a quart of water into a river that will flood.) But according to Broome, your emissions do not inflict harm *only* by causing a significant increase, or even *any* increase, in cumulative GHG levels. Your emissions inflict harm by changing weather patterns. Indeed, if Broome is correct, then even a *decrease* in cumulative GHG levels can (and probably would) inflict harm on some people, although, for some people, it would probably also avert harm.

2.3 A Tension between the Chaos Argument and Broome's Defence of Offsetting

Unfortunately, the claim that your emissions inflict harm by affecting a chaotic system and altering weather patterns, and not simply by increasing cumulative GHG levels, is very difficult to square with the unstated assumption underlying Broome's defence of offsetting in *Climate Matters*. If the Chaos Argument is sound (specifically, if the first three premises are true), then your GHG emissions will probably cause harm even if you offset your emissions. When you drive your gas guzzler for fun, your emissions of GHG will create atmospheric disturbances that

[11] The qualification 'almost' is added because a zero probability of an event is consistent with the occurrence of that event.

will escalate to cause significant changes in large-scale meteorological events, some of which will (probably) result in harm that would not otherwise have occurred. Offsetting your emissions probably would not undo these effects. In fact, insofar as offsetting would also cause (sufficiently large) atmospheric disturbances, offsetting would probably also inflict harm that would not otherwise have occurred.

This does *not* mean that if the Chaos Argument is sound there is no morally relevant difference between increasing and decreasing emissions of GHG. The Chaos Argument is consistent with the claim that *other things being equal,* increasing cumulative GHG emissions increases *the expectation* of total harm from climate change while decreasing cumulative GHG emissions decreases this expectation. Earlier we considered an analogy involving people trapped in a water tank to illustrate Broome's unstated assumption in *Climate Matters* that your emissions do harm only by increasing cumulative emissions. The Chaos Argument calls for a modification of the analogy. Suppose that the potential victims trapped in the tank are currently able to evade death by keeping their heads just barely above water. And suppose that the addition or removal of water from the tank creates waves that spread around the tank. The waves have high peaks and low troughs. The peaks go above the heads of some people and cause them to drown, while the troughs go below the heads of others, allowing them to breathe when they otherwise would not have been able to, and thus, avoid drowning. Adding or removing water from the tank changes how the waves interact, and thus changes the distribution of the peaks and troughs. It therefore alters the distribution of harm. But the expectation of *total* harm still increases with the water level; adding water still imposes a greater risk of harm than removing water, and the more water is added, the greater the risk of drowning for each person. What one cannot reasonably claim is that adding some water and removing the same amount of water results in "no harm".

If the Chaos Argument is sound, then Individual Denialism is almost certainly false, but so is Broome's claim that if you offset your emissions you do no harm by emissions. The Chaos Argument undermines Individual Denialism, but it also undermines Broome's defence of offsetting. How should Broome respond?

3. Harm Avoidance, Risk Avoidance, and Offsetting

In this section, I consider how Broome might respond to the tension between his argument against Individual Denialism and his defence of offsetting in *Climate Matters*. It seems very difficult to maintain both positions. I therefore suggest an alternative defence of the claim that you can satisfy your duty to limit your carbon footprint by offsetting your emissions, where this alternative defence does not depend on the claim that you have a duty to avoid doing harm by emissions. Whether such a defence is plausible *all things considered* is a question for another paper. This section motivates the alternative defence of offsetting and considers some potential worries that a more thorough defence would need to address.

In *Climate Matters* Broome claims that your moral duty vis-à-vis limiting your carbon footprint is a duty to avoid doing harm. But one possibility is that if the Chaos Argument is sound, such a duty is practically impossible to meet. This would be true if it were practically impossible to avoid causing "sufficiently large" atmospheric disturbances of the kind that escalate to a global scale and that have "literally zero probability" of doing no harm. For example, if "sufficiently large" atmospheric disturbances are not much larger than the flapping of a butterfly's wings, then it is unclear how you could realistically satisfy your duty of harm avoidance.[12]

If doing climate harm is practically unavoidable, then it makes little sense to claim that your duty to reduce your GHG emissions derives from a duty to "do no harm". A better option would be to focus on limiting the *risk* of harm that you impose by increasing emissions of GHG. If, as Broome and others have argued, the expectation of climate harm is an increasing function of cumulative GHG emissions, then *other things being equal* increasing GHG emissions imposes a greater risk of harm on the global population than reducing GHG emissions.[13] Broome might therefore claim that your individual duty to reduce your carbon footprint derives not from a duty of harm avoidance but rather from a duty of risk avoidance.

Whether offsetting your emissions would be sufficient to satisfy your individual duty of risk avoidance depends on the specific content of that duty. One possibility is that the duty is simply to ensure that your lifetime activity is at least *risk neutral* with respect to emissions of GHG. That your lifetime activity is risk neutral in this sense does not mean that "you do no harm by emissions". Rather, it means that in expectation, the climate harm that you inflict (and avert) by adding GHG to the atmosphere and the climate harm that you avert (and inflict) by subtracting GHG balance out. If your individual duty of risk avoidance is understood in this way then it seems you could, in principle, satisfy the duty by offsetting your emissions.

One potential worry about understanding your duty of risk avoidance in this way is that balancing the expected harm and the expected benefit (i.e. harm averted) by one's behavior often seems insufficient to satisfy one's moral duty vis-à-vis risk-imposition. For instance, it may seem implausible that one can justify imposing a serious risk on some person on the grounds that one's behavior also averts an equally serious risk to some other person, assuming that one has the option of not

[12] Ending your own life would not be sufficient, since even this would have knock-on effects of the kind that would escalate to affect weather patterns.

[13] Others who have argued that increasing GHG emissions increases the expectation of harm include Hiller (2011) and Nolt (2011).

imposing any risk at all.[14] A defender of offsetting might respond that if the Chaos Argument is sound, then one may not have the option of imposing *no risk* whatsoever. The defender of offsetting might also claim that if the Chaos Argument is sound, then the case of risk-imposition by emissions of GHG is different from a case in which you impose a risk on one person and avert an equally serious risk to someone else. The difference might be that if the Chaos Argument is sound, and if you choose to offset your emissions, then each unidentified person in the global population who might incur harm because of your choice might also avoid harm because of your choice. Suppose that if you were to offset your emissions, the expectation of harm associated with your offsetting would be the same as the expectation of benefit (i.e. harm averted) for each unidentified person in the global population.[15] Then with respect to each unidentified person, your choice of offsetting would be risk neutral. If this were true, then perhaps your choice to offset your emissions could be justified to each individual *ex ante*.

However, there is at least one complication. Offsetting is not sufficient to guarantee risk-neutrality with respect to each unidentified person. Offsetting means only that for any amount of GHG you add to the atmosphere, you subtract at least that much. It is possible to add and subtract GHG in such a way that some people incur *uncompensated* risk of harm. For example, suppose you emit 800 tons of GHG throughout your life and that you offset by giving money to an organization that plants trees. Thanks to your donation, the organization successfully reduces GHG emissions by 800 tons. But suppose that the emissions reduction target is achieved only several decades after your donation is secured (perhaps because it takes several decades for enough trees to be planted). Then many individuals who are subjected to risk by your emissions of GHG will not be alive in several decades (e.g. because they are already near the end of their natural lifespans). These individuals cannot benefit, even in expectation, from your offsetting. Because you subject these individuals to risk by your emissions, they may have complaints against you that would not be addressed by pointing out that, overall, your behavior was neutral regarding the expectation of total climate harm. A fully-developed version of the defence of offsetting that I am suggesting will need to say something about such cases.

[14] Hyams and Fawcett (2013, p. 95) raise a similar objection, although theirs involves a single victim and is not framed in terms of risk. They find it unacceptable that you would fulfil your duty to avoid inflicting harm for the reason that, although you did in fact harm someone, you also prevented an equally serious harm to the same person.

[15] This might include future people, not just those who exist now.

4. Conclusion

I have argued that Broome's claim that offsetting is sufficient to satisfy your duty to avoid doing harm by emissions is in tension with his refutation of Individual Denialism (ID). The refutation of ID is given by the Chaos Argument, but if this argument is sound, then the claim that offsetting is sufficient to do no harm by emissions is probably false. I suggested an alternative defence of the claim that you can satisfy your duty to limit your carbon footprint by offsetting your emissions, one that assumes your duty to limit your carbon footprint derives from a duty of risk avoidance. Your duty might be to ensure that your lifetime activity is at least risk neutral with respect to emissions of GHG. (Of course, there are other possibilities.) I considered the worry that balancing the expected harm and the expected benefit of your behavior may be insufficient to satisfy a duty of risk-avoidance. I also considered the possibility that if you offset, some individuals may incur uncompensated risks. A fully-developed version of the defence that I have suggested will need to address such issues. Yet, the shift to thinking about individual climate-related duties in terms of risk-avoidance seems warranted insofar as one accepts the Chaos Argument.

References

Broome, John. 2012. *Climate Matters: Ethics in a Warming World.* New York: W. W. Norton & Co.

Broome, John. 2019. "Against Denialism." *The Monist* 102: 110–129.

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

Hiller, Avram. 2011. "Climate Change and Individual Responsibility." *The Monist* 94: 349–68.

Hyams, Keith and Tina Fawcett. 2013. "The Ethics of Carbon Offsetting." *Wiley Interdisciplinary Reviews: Climate Change* 4, 91–98.

Maltais, Aaron 2013. "Radically Non-Ideal Climate Politics and the Obligation to at Least Vote Green," Environmental Values 22: 589–608.

Nolt, John. 2011. "How Harmful Are the Average American's Greenhouse Gas Emissions?" *Ethics, Policy & Environment,* 14:1, 3–10.

Sinnott-Armstrong, Walter. 2005. “It’s not my fault: global warming and individual moral obligations”, in *Perspectives on Climate Change: Science, Economics, Politics, Ethics. Advances in the Economics of Environmental Research,* Volume 5, edited by Walter Sinnott-Armstrong and R. B. Howarth, Elsevier, pp. 293–315.

Paul Bowman[1]

The Relevance of Motivations to Wrongdoing for Contributing to Climate Change[2]

This paper makes progress towards an account of moral wrongdoing for individual contributions to collectively-caused harms and substantial risks of harm, like the harms and risks stemming from climate change. To do so, this paper argues that an agent's motivations can be relevant to whether an agent's contribution to a collectively-caused harm or risk is morally wrong. Specifically, this paper argues that an agent's contribution to a collectively-caused harm or risk can be wrong in virtue of her motivations even when she does not intend to contribute to the harm or risk, but rather contributes to the harm or risk as a foreseeable but unintended side-effect of her otherwise good end.

[1] Institute for Futures Studies, paul.bowman@iffs.se.

[2] Financial support by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences) is gratefully acknowledged.

1. Introduction

In this paper, I make progress towards an account of moral wrongdoing for individual contributions to collectively-caused harms and substantial risks of harm. My aim in developing such an account is to evaluate, in a larger work, whether and to what extent ordinary individuals have acted wrongly by performing ordinary greenhouse gas-emitting activities. Ordinary greenhouse gas-emitting activities (henceforth *emitting activities*) include both activities that directly produce greenhouse gas emissions, like driving a car, as well as activities, like purchasing consumer goods and travelling by airplane, that indirectly produce greenhouse gas emissions by incentivizing others to produce (directly) greenhouse gas emissions.

My aim is to provide an account of moral wrongdoing that can help evaluate individuals' past emitting behavior, as opposed to one that (directly) provides advice on what individuals should do now and in the future. While questions of retrospective and prospective evaluations of behavior are obviously closely related, they can also come apart in important ways. My ultimate goal in evaluating individuals' past emitting behavior is to evaluate whether and to what extent these individuals have incurred duties of corrective justice to address climate change in virtue of this behavior. Fulfilling one's duty of corrective justice to address climate change may require different behavior than what one ought to have done to avoid incurring the duty in the first place.

My goal (in the larger work) is to evaluate whether and to what extent individuals acted *all-things-considered* wrongly by performing emitting activities. All-things-considered wrongness is here contrasted with *pro tanto* wrongness. An agent's act is *pro tanto* wrong when the act has at least one wrong-making feature that is not defeated by other normative features of the act.[3] An agent's act is all-things-considered wrong when the act is *pro tanto* wrong and the agent does not have an overriding (or outweighing) justification to perform the act.[4] So, for example, if A breaks B's finger merely because B uses the improper fork at a dinner party, then A acts *pro tanto* wrongly in virtue of wronging B by infringing B's right against bodily harm. Because A also lacks an overriding justification for wronging B in this way, A acts all-thing-considered wrongly. By contrast, if A breaks B's finger as an unavoidable side-effect of preventing three other persons from having their legs broken, then it is plausible that A acts *pro tanto* wrongly but not all-things-considered wrongly. A acts *pro tanto*

[3] Liability justifications often serve to defeat wrong-making features of acts. For example, while it is typically wrong for me to kill a person, if that person has made himself liable to be killed by me—by, say, culpably threatening me with lethal force, then my killing him does not wrong him, and so is not even pro tanto wrong.

[4] Lesser evil justifications are typically overriding, rather than defeating.

wrongly because A still wrongs B by infringing B's right, but A is nevertheless justified in doing so, given the much greater harm that A prevents. Henceforth, 'wrong' (and cognates) will refer to all-things-considered wrongness, unless otherwise indicated.

The account of moral wrongdoing for individual contributions to collectively-caused harms and substantial risks of harm that I defend in what follows centers on the relevance to wrongdoing of an agent's motivations with which the agent contributes to the harm or risk. According to my understanding, an agent's motivations with which the agent performs some act consists in "the whole mental process that results in the action" (Wedgewood, 2011). Motivations include both intentions, which directly control one's behavior, as well as mental states like desires, beliefs, dispositions, and emotions, which indirectly control one's behavior by leading one to form certain intentions (Wedgewood, 2011).

When I say that an agent's motivations are relevant to wrongdoing, I mean that the following principle is both true and relevant in a wide range of cases:

> the *Motivations Principle*: An act performed with certain motivations can be more seriously *pro tanto* morally wrong than an act performed with different motivations, other things being equal.

If some act A_1 is more seriously *pro tanto* morally wrong than some alternative act A_2, then A_1 is harder to justify than A_2, other things being equal. Moreover, one might be liable to bear greater compensatory or punitive costs in virtue of performing A_1 rather than A_2, other things being equal.[5] Note that saying that A_1 is more seriously *pro tanto* wrong than A_2 does not imply that A_2 is *pro tanto* wrong. It does imply that A_1 is at least *pro tanto* wrong, though A_1 may be all-things-considered permissible.

Note also that the Motivations Principle makes a claim about the (relative) moral status of *acts*, rather than about, e.g., the goodness or badness of the characters of the agents who perform the acts or the goodness or badness of the states of affairs in which the acts are performed. It is uncontroversial that motivations can be relevant to the evaluation of an agent's character, at least. It is far more controversial whether motivations can be relevant to the evaluation of acts.[6]

Finally, the 'other things being equal clause' is crucial. The Motivations Principle

[5] Though whether someone is liable to bear any such costs might depend on whether the act is all-things-considered wrong.

[6] Among philosophers who argue that motivations are relevant to wrongdoing include Quinn (1989), McMahan (2009a, 2009b), Tadros (2011), and Liao (2012). Philosophers who have argued that motivations are irrelevant to wrongdoing include Thomson (1991, 1999), Kamm (2007), Scanlon (2008), and Norcross (1991).

holds that one act can be more seriously *pro tanto* wrong than another act holding everything constant except for the motivations with which the agents perform the respective acts (e.g., the acts are performed with the same bodily movements and have the same actual and expected consequences).

Although the Motivations Principle is controversial among moral philosophers, I will argue that the principle is both true and relevant in many cases in which an agent contributes to a collectively-caused harm or substantial risk of harm. In particular, I will argue that the Motivations Principle is relevant in cases that can help us evaluate whether and to what extent ordinary individuals acted wrongly by performing emitting activities.

The remainder of the paper is structured as follows. Section 2 briefly considers two accounts that seek to explain why performing an emitting activity can be wrong. My discussion of these accounts will help motivate the importance of considering an agent's motivations when evaluating the moral status of contributions to collectively-caused harms and risks. Section 3 provides support for the relevance of the Motivations Principle in some cases featuring collectively-caused harms and risks, though these are cases in which the motivations of the contributing agents are very different from those with which individuals typically perform emitting activities. Section 4 argues that the Motivations Principle is relevant in cases in which the agents' motivations more closely align with those of individuals who perform emitting activities. In particular, it argues that there can be morally relevant differences among the motivations of agents who contribute to a collectively-caused harm or risk as a side-effect of pursuing some good end. Section 5 briefly considers some additional questions concerning the relevance of motivations in such cases, including the relevance of self-interested versus benevolent motivations, as well as whether wrongdoing can depend on one's culpable ignorance. Section 6 briefly concludes.

2.

In this section, I briefly consider two accounts that seek to explain why performing an emitting activity can be wrong. My discussion of these accounts will help motivate the importance of taking into consideration an agent's motivations when evaluating the moral status of the agent's contribution to a collectively-caused harm or risk.

The accounts that I discuss in what follows are motivated in large part by the widely-accepted assumption that although anthropogenic climate change has already caused harm to persons and will, in the absence of significant and costly preventative action, cause harm to many more persons, it seems highly unlikely that an individual's performance of an emitting activity makes any difference to whether or the extent to which any actual climate change-induced harm has occurred or will occur.[7] Therefore, to explain why performing an emitting activity can be wrong, one cannot simply appeal to the harm that performing the activity would individually cause, in the sense that performing the activity would be necessary for the harm to occur.

Now even though it is *highly unlikely* that an individual's performance of an emitting activity makes a difference to whether or the extent to which any climate change-induced harm occurs, it seems implausible that the probability is *zero*. Accordingly, some philosophers argue that to explain why performing an emitting activity can be wrong, we should instead focus on the fact that performing an emitting activity increases the *risk* that some persons will suffer climate change-induced harm they would not otherwise suffer.[8] Thus, these philosophers endorse a principle like the:

> *Risk Principle:* An act is *pro tanto* morally wrong when the act individually increases the risk that a person will suffer harm, other things being equal; the extent to which the act is *pro tanto* morally wrong is determined by the extent to which the act individually increases the magnitude of expected harm, other things being equal.[9]

Given the Risk Principle, and if we assume that performing an emitting activity individually increases the risk that some persons will suffer climate change-induced harm that they otherwise would not suffer, then performing an emitting activity is *pro tanto* wrong, other things being equal. Moreover, if we assume that performing an emitting activity increases the magnitude of expected climate change-induced harm in proportion to the amount of emissions produced by the performance of the activity, then the extent to which performing the activity is *pro tanto* wrong is also proportional to the amount of emissions produced by the performance of the activity. So whether performing any particular emitting activity is all-things-considered wrong depends both on the amount of emissions produced by the performance of the activity, as well as on the strength of the reasons the agent has

[7] See, e.g., Sinnott-Armstrong 2005, Sandberg 2011, and Cripps 2013. In his 2012, Broome seems to deny this assumption, although in his 2019 he defends the claim that an individual's emissions individually increase *expected* (climate change-induced) harm.

[8] See, e.g., Hiller 2011, Lawford-Smith 2016, and Broome 2019. Some consequentialists have appealed to expected *utility* to explain why contributing to climate change can be wrong. See, e.g., Morgan-Knapp and Goodman (2015).

[9] Roughly, the magnitude of expected harm is a weighted average of the magnitude of the harm in each possible outcome multiplied by the probability that the outcome will occur.

for performing the activity (e.g., determined by, e.g., the magnitude of the harm prevented or good produced by performing the activity).

It is not my goal to evaluate fully either the Risk Principle or the broader account of wrongdoing for performing an emitting activity that relies on it.[10] I will, however, briefly note three potential problems with the account, some of which have been noted by others. First, it is not obvious that it is always *pro tanto* wrong to increase the risk that a person will suffer harm.[11] Virtually everything one does increases risks of harm to others. For example, when I enter my apartment, I (ever so slightly) increase the probability that the floor beneath me will collapse on my neighbors, injuring them. But it does not seem to be even *pro tanto* wrong for me to enter my apartment. Although I do not think that this problem is insurmountable, it indicates that the Risk Principle needs either additional defense or modification.

Second, and relatedly, it is unclear to what extent performing an emitting activity individually increases the magnitude of expected climate change-induced harm. Although I think that we can confidently state that performing an emitting activity individually increases the magnitude of expected climate change-induced harm to *some* degree, there may be reason to think that the extent that performing an emitting activity individually increases the magnitude of expected climate change-induced harm is so small that the *pro tanto* wrongness of performing the activity will be outweighed by virtually any normative reason in favor of performing the activity.[12] There is obviously much more that can be said here, but my aim is simply to note the potential problem.

Third, and most importantly for my purposes in this paper, it seems that whether and to what extent it is *pro tanto* wrong to contribute to a collectively-caused harm or substantial risk of harm (like the harms and risks stemming from climate change) does not depend *solely* on whether and to what extent the contribution individually increases the magnitude of expected harm. This is not a problem for the Risk Principle as such, but it does show that a complete account of why it can be wrong to perform an emitting activity likely has to go beyond the Risk Principle.

[10] For a recent extended defense of this basic account, see Broome 2019.

[11] Versions of this objection have been put forward by Sinnott-Armstrong 2005, Posner and Weisbach 2010, and Jamieson 2014 among others. For potential solutions to this problem, see Hiller 2011 and Lawford Smith 2016.

[12] There may be multiple factors that reduce the extent that performing an emitting activity otherwise individually increases the magnitude of expected climate change-induced harm. For example, one's performing an emitting activity may have certain (expected) market effects (e.g., increasing the expected price of oil, thereby decreasing expected aggregate demand). See Hale 2011 for an interesting discussion of some relevant issues.

To illustrate, consider the following case:

> *Unnecessary Contribution:* Each of three agents—A, B, and C—independently dumps a bucket of the same kind of chemical waste from his malfunctioning heating unit (which is far too expensive to replace at this time) into a creek that feeds into V's water supply. The wastes completely mix before reaching V's water supply. Each dumps the waste in the creek to avoid an otherwise unavoidable moderately expensive disposal cost. At the time of acting, each agent knows for certain that (a) if one or fewer buckets of waste is dumped, then V will not suffer any harm at all, but (b) if more than one bucket of waste is dumped, then V will go permanently blind. Each agent also knows for certain that no one else is at risk of being harmed, and that V will not suffer a harm greater than blindness, no matter how much waste is dumped. Finally, each agent knows for certain that the other two agents will dump their waste close to the same time that he does and that he cannot affect what the other agents do. V drinks the waste-laced water and goes permanently blind.

It is highly plausible that each agent acts seriously wrongly by contributing to V's blindness, even though, at the time of acting, no agent's contribution individually increases the risk of harm to V, given that each knows for certain that V would go blind even if he does not dump the waste. (Henceforth, I will call a contribution to a collectively-caused harm or risk of harm that does not individually increase the magnitude of (expected) harm an *unnecessary contribution*).[13]

Although *Unnecessary Contribution* is not intended to be closely analogous to the circumstances in which individuals perform emitting activities, the case suggests that even if performing an emitting activity does not individually increase the magnitude of expected climate change-induced harm, or only increases it by a trivial degree, performing the activity may nevertheless be wrong. Conversely, if performing an emitting activity *does* increase the magnitude of expected climate change-induced harm by a non-trivial degree, then performing the activity may be wrong for other reasons as well, thereby contributing to the *overall* extent that performing the activity can be seriously wrong.

[13] Here and throughout, I focus on cases that might be best described as cases of *overdetermination*, i.e., cases in which more than one agent contributes to the harm, but a smaller number of these contributions would have been sufficient for the harm to occur. We can contrast these with *preemption* cases in which more than one agent contributes to the harm, but a smaller number of these contributions would not have been sufficient for the harm to occur; however, had one or more agents not contributed, then others would have provided the necessary contributions for the harm to occur. I think that there may be morally relevant differences in the moral status of contributions to these different kinds of case, but I cannot explore this issue here. At the very least, I think that some climate change-induced harms are overdetermined.

Let's consider, then, a second account that seeks to explain why it can be wrong to perform an emitting activity and that seems well-suited to accommodate the intuitive verdict in *Unnecessary Contribution*. This account centers on the:

> *Bare Contribution Principle:* An act is *pro tanto* morally wrong when the act contributes to a collectively-caused harm or substantial risk of harm; the extent to which the act is *pro tanto* wrong depends on (a) the extent of the contribution (independently of the extent to which the contribution individually increases the risk), and (b) the magnitude of the overall harm or expected harm to which the act contributes, other things being equal.[14]

Notice that the Bare Contribution Principle implies that contributing to a collectively-caused harm or risk is *pro tanto* wrong even if the contribution is unnecessary. Thus, the Bare Contribution Principle straightforwardly supports the conclusion that in *Unnecessary Contribution*, each agent acts at least *pro tanto* wrongly. Moreover, because (a) each agent makes a substantial contribution to the harm to V, (b) the harm to V is very large, and (c) each agent has, at best, only a very weak reason in favor of contributing to the harm (avoiding a moderate disposal cost), the Bare Contribution Principle plausibly supports the conclusion that each agent acts all-things-considered wrongly.

In addition, the Bare Contribution Principle supports the conclusion that performing an emitting activity is at least *pro tanto* wrong, given that performing an emitting activity contributes to climate change-induced harms and substantial risks of harm, even if performing the activity does not individually increase the magnitude of climate change-induced (expected) harm. According to the Bare Contribution Principle, whether any particular emitting activity is all-things-considered wrong depends on the overall magnitude of (expected) climate change-induced harm to which one's performance of the activity contributes, the amount of greenhouse gas emissions the performance of the activity produces, and the strength of the reasons the agent has for performing the activity.

Notice, however, that the strength of the duty, given by the Bare Contribution Principle, to avoid contributing to a collectively-caused harm or risk seems to be relatively weak, at least in comparison to the duty given by the Risk Principle, other things being equal.[15] For example, in *Unnecessary Contribution*, it seems that if one of the agents could, by dumping his waste, prevent even a modest harm to a different

[14] The Bare Contribution Principle is similar to a principle endorsed by Barry and Overland (2016). Cripps (2013) also seems to endorse a principle like the Bare Contribution Principle.

[15] A similar point is made by Barry and Overland (2016).

person (a bad stomach ache, say), then he might be permitted to do so given that he knows for certain that V will suffer the same harm regardless. But if his dumping the waste would individually increase the risk that a person would go blind by even a small amount, then it seems that he would need a much stronger reason to dump his waste than preventing a stomach ache.

Yet there is a bit of a puzzle here. It seems that each of the agents in *Unnecessary Contribution* act in a way that is *seriously* wrong. For example, it seems that each of the agents is liable to severe punishment for dumping his waste. And yet, as I have said, the duty given by the Bare Contribution Principle appears to be relatively weak. It is puzzling how three agents could act seriously wrongly by contributing to an innocent person going blind but each of them violates only a weak duty to refrain from contributing to the harm.

In the next section, I will argue that the Bare Contribution Principle is false. An act that contributes to a collectively-caused harm or substantial risk of harm is not sufficient to make the act *pro tanto* wrong. Nevertheless, I think that the Bare Contribution Principle is close to the truth. I will argue that whether and the extent to which making an unnecessary contribution to a collectively-caused harm or risk is *pro tanto* wrong can depend on the motivations with which the agent makes the contribution.[16] Showing that, and how, motivations are relevant to wrongdoing in these cases will also help resolve the puzzle I sketched above.

3.

In this section, I argue for the

> *Motivations Principle*: An act performed with certain motivations can be more seriously *pro tanto* morally wrong than an act performed with different motivations, other things being equal.

More specifically, I argue that the Motivations Principle is relevant in cases featuring unnecessary contributions to collectively-caused harms and risks of harm. To do so, I consider a pair of cases, first presented and discussed by Victor Tadros (2011), in which it is highly plausible that the agents' motivations are relevant to the moral status of their acts. I then defend the relevance of the Motivations Principle in these cases against those who reject the principle.

[16] Barry and Overland (2016) argue that the stringency of the duty to avoid making an unnecessary contribution to a collectively-caused harm depends on the motivations with which one acts, but their arguments are very different than the ones presented in what follows.

It should be noted at the outset that the behavior represented in these cases is not intended to be analogous to most individuals' emitting behavior. The cases are primarily intended to establish the broader relevance of the Motivations Principle in cases featuring unnecessary contributions to collectively-caused harms and risks. Additionally, as we shall see, one of the cases serves as a counterexample to the Bare Contribution Principle. In subsequent sections, I will present and discuss cases that represent behavior that is more closely analogous to the emitting behavior of many individuals (at least in some relevant respects), and in which, I argue, the Motivations Principle is also relevant.

As I noted above, the Motivations Principle is controversial among moral philosophers. Philosophers who accept the principle argue that it can help explain a number of widely-held judgments in a range of cases. Consider, for example, the following familiar pair of cases:

> *Strategic Bomber:* A pilot in a just war bombs a munitions factory intending to destroy its productive capacity in order to induce the enemy to surrender. The pilot foresees that bombing the factory will kill some number of innocent civilians who live nearby.

> *Terror Bomber* A pilot in a just war intends to kill some number of innocent civilians living near a munitions factory in order to induce the enemy to surrender. The pilot bombs the munitions factory because bombing this location will ensure that the intended number of innocent civilians are killed.

Suppose that the strategic bomber and the terror bomber perform exactly the same physical movements with exactly the same actual and expected consequences. It is plausible that there are some such circumstances in which the strategic bomber's act would be permissible while the terror bomber's act would be impermissible (e.g., the respective bombings would save 1000 innocent lives by ending the war but kill 100 innocents living nearby). The Motivations Principle appears to help explain this judgment—the bombers act with different motivations. Whereas the strategic bomber kills the civilians as a foreseen but unintended side-effect of destroying the munitions building to end the war, the terror bomber kills the innocent civilians as an intended means of ending the war.

More central to my purposes in this paper, the Motivations Principle appears to be relevant in some cases featuring unnecessary contributions to collectively-caused harms and risks of harm. Tadros (2011) presents a pair of such cases.

Consider, first, the following case, which is adapted from Tadros (2011, p. 159):[17]

> *Intentional Contribution:* A mob boss offers a substantial reward for killing V. Each of three agents—M, N, and P—intends to kill V by pouring a vial of lethal poison into V's water supply in order to earn the reward. Each agent knows for certain that each of the other agents will pour her vial close to the same time that she does, and that the poisons will mix before reaching V. Each agent knows that one vial of poison is sufficient to kill V. Each agent also knows that if either one or two vials of poison are poured into V's water supply, V will suffer a very long and agonizing death, but if three or more vials are poured, then V will die very quickly and painlessly. While each agent knows that she cannot affect what any other agent does, each would strongly prefer that no other agent pours their vial (because each wants the reward entirely to herself). Each agent pours her vial of poison into V's water supply, and V dies a quick and painless death as a result.

It is highly plausible that each agent acts seriously wrongly by pouring her vial of poison into V's water supply, even though no agent's act is necessary for V's death, and even though had any one agent refrained from pouring her vial, V would have been significantly worse off as a result.

Yet now consider the following variation (also adapted from Tadros 2011, p. 159):

> *Palliative Contribution:* The same set-up as *Intentional Contribution,* but the third agent is Q (*rather* than P). While each M and N pours her vial in order to kill V (to collect the reward, as before), Q pours her vial solely to reduce V's suffering. Q would not have poured her vial if doing so would not have reduced V's suffering, and Q will not collect the reward.[18]

It is intuitively plausible that Q's act is morally permissible, even though Q contributes to V's death.

Compare, then, P's act in *Intentional Contribution* to Q's act in *Palliative Contribution*. Intuitively, P acts wrongly whereas Q acts permissibly. Yet the agents perform the same physical movements with the same actual and expected consequences. The only difference between P's behavior and Q's behavior is the motivation with which each agent performs her act. P is motivated to kill V as a means of

[17] Much of my subsequent discussion of these cases builds on Tadros' important, though somewhat brief, discussions in his 2011 and 2013.

[18] I assume that V is unable to consent to Q's choice to pour the poison.

earning the reward, whereas Q is motivated to reduce V's suffering.[19] It is plausible, then, that the Motivations Principle is relevant in at least some cases featuring unnecessary contributions to a collectively-caused harm (Tadros, 2011).

Those who reject the Motivations Principle typically claim that although an agent's motivations are not relevant to the moral status of the agent's act, an agent's motivations are often relevant to the evaluation of the agent's moral *character* (how morally good or bad the agent is) and, perhaps, how morally *blameworthy* the agent is.[20] These philosophers would therefore respond to the pair of cases presented above by claiming that although we may naively believe that our intuitive judgments are tracking differences in the moral status of the acts of P and Q, our judgments are really tracking differences in the agents' moral characters or moral blameworthiness.[21] So these philosophers would claim that the pair of cases give us strong evidence that P is a bad person and that Q is a good person, or that P but not Q is liable to blame. But regarding the moral status of the acts themselves, either both acts are morally permissible or both are morally impermissible.

One problem with this position is that it is genuinely unclear whether we should say that P's act and Q's act are both permissible or that both are impermissible (cf. Tadros, 2013). Are both acts wrong because both contribute to killing an innocent person, or are both permissible because both save an innocent person from suffering a horribly painful death? Neither option is appealing.

Suppose that both acts are morally permissible. Given that in *Intentional Contribution*, M, N, and P are identically situated, each of their acts would be morally permissible. This implies that an innocent person is killed by three agents acting independently, each of whom intends to kill V for money, and yet no agent does anything wrong. This implication is highly counterintuitive. Moreover, if wrongdoing is necessary for liability to pay compensation (as seems plausible), then none of the agents would owe compensation to V's family.[22] If wrongdoing is necessary

[19] There is perhaps a question concerning whether Q intends to *kill* V in order to reduce V's suffering. I'm not sure (perhaps we can describe the case in different ways), but I do not think that it matters morally, given that Q's sole, *ultimate* aim is to reduce V's suffering.

[20] See, for example, Thomson 1999, Scanlon 2008, and Norcross 1991. I should note that those who wish to hold that an agent's motivations are relevant to their blameworthiness must specify for *what* the agent is blameworthy. It seems implausible that an agent could be blameworthy for performing a permissible act. Therefore, it seems that the agent would have to be blameworthy either for their bad motivations (e.g., bad intentions or desires), or perhaps for acting on their bad motivations (if indeed this can be distinguished from blameworthiness for performing the act itself).

[21] See esp. Norcross (1991) who argues that those who judge that intentions (or motivations) are relevant to wrongdoing are mistaken as to what their judgments are tracking.

[22] Perhaps one could say that each agent acts *pro tanto* wrongly, but not all-things-considered wrongly. Since it is plausible that an agent can be liable to pay compensation on the basis of acting merely *pro tanto* wrongly, this would allow one to say that each can be liable to compensate V's family. First, this response does not address the question of liability to punishment, which plausibly requires all-things-

for punishment (as seems even more plausible), then none of the agents would be liable to punishment. Again, these implications are highly counterintuitive.

Suppose instead that the acts of P and Q are both morally impermissible. This would mean that Q acts wrongly even though Q intentionally saves V from suffering a horrible death. Moreover, because Q is a fully responsible moral agent (we can assume) who knows all of the relevant non-moral facts about the situation, presumably Q is also *culpable* for wrongly contributing to V's death—that is, Q does not have either justification or excuse for her wrongful act. Insofar as she is culpable for wrongly contributing to V's death, presumably Q would be liable both to compensation and to punishment. These implications are counterintuitive.

Yet if the Motivations Principle is true, then we can provide the following straightforward and compelling evaluation of P and Q's behavior. In *Palliative Contribution,* Q's act is morally permissible, and Q is morally praiseworthy for saving V from horrible suffering (despite the fact that Q's act also contributes to V's death). By contrast, in *Intentional Contribution,* P's act is morally impermissible, and P is culpable for wrongly contributing to V's death (despite the fact that P's act also saves V from horrible suffering). Moreover, because P (but not Q) is culpable for wrongly contributing to V's death, P (but not Q) is liable both to compensation and to punishment.

Consider now the question of whether Q's act is even *pro tanto* wrong. It seems to me that Q's act does not wrong V and so is not *pro tanto* wrong. Q's situation seems to be relevantly similar to a situation in which an agent harms a person in a respect in order to prevent a worse harm to that person. In these cases, the agent typically does not act even *pro tanto* wrongly. Suppose, for example, that I cut off your leg because if I do not, then you will die from infection. Although I harm you in a respect by cutting off your leg, I do not wrong you, given that cutting off your leg prevents a greater harm (this assumes, of course, that I am motivated to prevent this greater harm and not, for example, to cause you great agony).[23] Similarly, we should say that Q does not act even *pro tanto* wrongly by pouring her poison.

If I am correct, then despite contributing to a collectively-caused harm, Q does not act even *pro tanto* wrongly. Thus, *Palliative Contribution* is a counterexample to the Bare Contribution Principle. That an act contributes to a harm or substantial

considered wrongdoing. Second, it seems odd to hold that a person is harmed by an action in a way that wrongs the person, but that the person is simultaneously benefited by the action in a way that makes the wronging permissible. It is of course common for a permissible act to both harm someone in a respect but benefit the person overall (e.g., cutting off one's leg to stop the spread of infection). But typically in such cases, it does not seem that the act wrongs the person. It does not, for instance, seem that compensation is owed in these cases. See my discussion of whether Q acts *pro tanto* wrongly, below.

[23] This judgment also presupposes that the person either consents to the procedure or is unable to give consent.

risk of harm is not sufficient to make the act even *pro tanto* wrong. Whether a contribution to a harm or risk is *pro tanto* wrong can depend on the motivations with which the agent makes the contribution.

It is important to say something about why motivations are relevant to wrongdoing in *Intentional Contribution* and *Palliative Contribution.* Consider the following, familiar account of the relevance of motivations to wrongdoing, which we can call *the respect account.*[24] According to the respect account, acting with certain motivations can be tantamount to disrespecting a being who, in virtue of the being's high moral status (their *personhood*), deserves respect. Given that agents have a strong moral reason to respect beings who deserve respect, that one's act disrespects a person is a significant wrong-making feature of the act. Therefore, acting with certain motivations can be a significant wrong-making feature of one's act.

The respect account is compatible with different theories concerning which motivations are morally significant to which acts. It is plausible, and also widely held among philosophers who endorse the respect account, that an agent can disrespect a person by harming the person as an intended means of fulfilling the agent's ends.[25] Typically, to harm a person as an intended means of fulfilling one's ends is to regard and treat that person as an object, available for use by the agent, rather than as a being who is an end-in-herself.[26] For example, when the terror bomber kills the innocent civilians as a means of ending the war, the bomber arguably regards and treats these civilians as objects and not as persons with ends of their own. It is plausible, then, that harming a person as an intended means of fulfilling one's ends can be more seriously *pro tanto* wrong than other forms of harming, other things being equal.[27]

The respect account, together with the claim that it is disrespectful to harm a person as an intended means of fulfilling one's ends, can accommodate the judgment that in *Intentional Contribution,* P acts wrongly by pouring the vial of poison into V's water supply, while in *Palliative Contribution,* Q acts permissibly by doing the same. P contributes to killing V as an intended means of fulfilling P's end of securing the reward. Therefore, P acts in a way that is tantamount to disrespecting V, which is seriously *pro tanto* wrong. By contrast, even though Q contributes to V's

[24] This basic account is endorsed, in some form, by Quinn (1989), Liao (2012), Nelkin and Rickless (2014) among many others. Tadros (2011, Ch. 7) also appears to endorse a version of the account. My presentation of this account draws on Liao's (2012) presentation of it. Liao attributes the account to Kant and Quinn.

[25] E.g., Liao 2012, Quinn 1989, Nelkin and Rickless 2014, Tadros 2011.

[26] Liao 2012, cf Quinn 1989.

[27] This claim is captured, more formally, by the well-known Doctrine of Double Effect, which (in one of its formulations) holds that harming as an intended means of fulfilling one's ends is more seriously *pro tanto* wrong than harming as a foreseen but unintended side-effect of pursuing one's ends, other things being equal.

death, Q is motivated solely to benefit V. Q does not regard or treat V as a means of fulfilling Q's end. Rather, Q regards and treats V as an end-in-herself, as a being who deserves respect. So although Q contributes to V's death, Q does not act in a way that is tantamount to disrespecting V. Therefore, Q's act lacks a significant wrong-making feature that P's act has.

Note again that neither P's act nor Q's act is (or is intended to be) analogous to the performance of emitting activities by ordinary individuals. Virtually no one, I assume, performs emitting activities intending to cause climate change-induced harm as a means of fulfilling their ends. Moreover, relatively few people perform emitting activities intending to ameliorate the climate change-induced harms that others have suffered or are at risk of suffering. Rather, it seems that most individuals who perform emitting activities are motivated to benefit themselves or those specially related to them, and contribute to climate change-induced harms and risks as foreseen (or at least foreseeable) side-effects of pursuing these ends.

In the next section, I will discuss cases that are, in some important respects, more analogous to the performance of emitting activities by ordinary individuals. More specifically, I will argue that there are morally relevant differences among the motivations of agents who contribute to a collectively-caused harm or risk as a foreseen but unintended side-effect of pursuing their ends. I will also extend the respect account to explain why these differences are morally relevant.

4.

To begin examining the morally relevant differences among the motivations of agents who contribute to a collectively-caused harm or risk as a foreseen but unintended side-effect of pursuing their ends, consider the following case. The reader will notice that this case is *Unnecessary Contribution* from Section 2 but with a few additional details.

> *Indifferent Contribution (Self-Interest)*: Each of three agents—A, B, and C—independently dumps a bucket of the same kind of chemical waste from his malfunctioning heating unit (which is far too expensive to replace at this time) into a creek that feeds into V's water supply. The wastes completely mix before reaching V's water supply. Each dumps the waste in the creek to avoid an otherwise unavoidable moderately expensive disposal cost. At the time of acting, each agent knows for certain that (a) if one or fewer buckets of waste is dumped, then V will not suffer any harm at all, but (b) if more than one bucket of waste is dumped, then V will go permanently blind. Each agent knows for certain that no one else is at risk of being harmed, and that V will not suffer a harm greater than

> blindness, no matter how much waste is dumped. Finally, each agent knows for certain that the other two agents will dump their waste close to the same time that he does and that he cannot affect what the other agents do. Each agent dumps his waste solely to avoid the disposal cost, and each agent would have dumped his waste to avoid the cost even had his contribution been individually necessary for V's blindness.[28] For example, each would have dumped his waste even had one (or both) of the other agents failed to dump his waste. V drinks the water and goes permanently blind.

Each agent makes an unnecessary contribution to V's blindness as a foreseen but unintended side-effect of benefiting himself (avoiding the disposal cost). Nevertheless, it is highly plausible that each agent acts wrongly. Here is a short argument for that conclusion. Given that all agents perform the same act with the same actual and expected consequences and the same motivations, either all of the agents act wrongly or all act permissibly. It is implausible that all act permissibly. This is because it is implausible that three agents can act in ways that jointly cause a person to go blind while generating only modest benefits for themselves but no agent acts wrongly. Therefore, each agent acts wrongly.

In this case, what makes each of their acts wrong is not, or is not merely, that each is motivated by self-interest, though it is at least plausible that this is relevant to the extent to which the act is wrong. I will say more about the relevance of self-interest below. For now, it will be helpful to introduce and consider a case similar to *Indifferent Contribution (Self-Interest)*, but where each agent is instead motivated by benevolence. Doing so will help focus our attention on a feature that is common to the agents' motivations in both variations and that is morally significant.

Consider, then, a variation of the case in which three agents (G, H, and J) each dumps his waste not to avoid the disposal cost, but solely in order to soothe a stranger's very painful stomach ache (each agent's dumping his waste soothes the stomach ache of exactly one stranger; three strangers in total have their stomach aches soothed).[29] As before, each agent would have dumped his waste to soothe the stomach ache even had his contribution been individually necessary for V's blindness. Call this variation *Indifferent Contribution (Benevolent)*. It is highly plausible that each agent acts wrongly in this variation as well. As before, it is hard to believe that three agents can act in ways that jointly cause a person to go perma-

[28] Assume, also, that each agent would know that his contribution is necessary for the harm.

[29] We can tell a story about how this is supposed to work. Suppose that each agent dumps his waste into the creek at a different point. Before mixing with the other waste, a tiny bit of the waste seeps into the stranger's water supply. The bit of waste mixes with a certain chemical in that person's pipes, and the resulting mixture has stomach-ache soothing effects. Yet enough of the agent's waste makes its way into V's water supply to mix with the other agents' waste, resulting in V's blindness.

nently blind while producing modest benefits for three different persons but no agent acts wrongly.

I suggest that in both variations of *Indifferent Contribution,* it is morally significant that each agent *would have* dumped his waste in pursuit of his end (either the self-interested one or the benevolent one) even in circumstances in which his contribution is necessary for the harm to occur. More formally, I suggest that the following principle is both true and relevant in both variations of *Indifferent Contribution*:

> *Counterfactual Independent Impermissibility Principle (CIIP)*: An act that causes or contributes to a collectively-caused harm or substantial risk of harm as a foreseen (or foreseeable) *side*-effect of pursuing some end can be more seriously *pro tanto* wrong when and to the extent that the agent would have performed the act in pursuit of the end in circumstances in which the act is independently impermissible, other things being equal.

Let me clarify a few aspects of CIIP before arguing for its relevance in *Indifferent Contribution* (*Benevolent*). (I will focus primarily on this variation since it does not have a potentially confounding factor—the self-interested motive—that the *Self-Interest* variation has).

First, an act is *independently impermissible* if the act is all-things-considered morally impermissible independently of the motivations with which the agent performs the act.[30] For example, an act is typically independently impermissible when the act foreseeably causes harm to an innocent person but does not prevent a much greater harm or produce a much greater good.[31]

Second, CIIP describes a respect in which motivations can be relevant to wrongdoing. This is because there is an important relationship between an agent's motivations with which she performs an act and what the agent would have done in

[30] I borrow the term "independently impermissible" from Walen 2006. Walen's "Doctrine of Illicit Intentions" has some overlap with my CIIP. Roughly, the Doctrine of Illicit Intentions holds that it is impermissible to form an intention that would direct one to perform independently impermissible acts in various circumstances, even if one only performs independently permissible acts. Walen denies, however, that forming such an intention is directly relevant to the wrongdoing of the (independently permissible) acts that one does perform. CIIP, by contrast, holds that acting with such an intention can be directly relevant to the wrongdoing of one's act.
CIIP also has some overlap with a proposal put forward by Pinkert (2015). Pinkert argues that act consequentialism should be supplemented with a principle that holds that agents ought to be such that they would act optimally in counterfactual scenarios. However, Pinkert interprets this principle as a requirement of moral virtue. So it seems to me that failing to meet this requirement is a defect in moral *character* (rather than a defect in one's act).

[31] Moreover, the terror bomber's bombing is independently impermissible in all and only the circumstances the strategic bomber's bombing is independently impermissible.

different circumstances. We can describe this relationship in at least two different ways, depending on one's underlying theory of action. First, we can say that, in general, an agent's intention with which she performs some act is very complex, directing an agent's behavior in pursuit of some end across a range of circumstances. So, for example, an agent's intention with which the agent performs an act A_1 to achieve some end E_1 in circumstances C_1 directs the agent to perform A_1 to achieve E_1 if C_1, (or C_2 or C_3) obtains, and to perform A_2 to achieve E_1 if C_4, C_5, or C_6 obtains, but to abandon E_1 if C_7, C_8, or C_9 obtains, etc.[32] Thus, an agent's intention with which the agent performs an act to achieve some end in some circumstances also determines which acts the agent would have performed to achieve that end had different circumstances obtained. An agent's underlying motivational mental states—her beliefs, desires, dispositions, etc.—determine the precise character of this complex intention.

Alternatively, we might say that an agent's intention directs the agent to perform an act for an end across a range of circumstances. So an agent's intention with which the agent performs an act A_1 to achieve some end E_1 in circumstances C_1 directs the agent to perform A_1 to achieve E_1 if C_1 (or C_2 or C_3) obtains. But had C_4, C_5, or C_6 obtained, the agent would have had a different intention directing the agent to perform A_2 to achieve E_1. Had C_7, C_8, or C_9 obtained, the agent would have had yet a different intention directing the agent to perform A_3 to achieve a different end E_2, and so on. On this picture, which intention an agent has, and hence which act the agent performs for which end in which circumstances, is determined by an agent's underlying motivational mental states—her beliefs, desires, etc. Moreover, which intention the agent would have had, and hence which acts the agent would have performed to achieve the same end had different circumstances obtained, is also determined by the agent's underlying motivational mental states.

To show that CIIP is relevant in *Indifferent Contribution (Benevolent),* let's take a closer look at the act of one of the agents, J. Recall that the case stipulates that J would have dumped the waste in order to prevent the stomach ache even in circumstances in which dumping the waste is necessary for V's going blind (e.g., circumstances in which exactly one of G or H dumps their waste). It is clearly impermissible, independently of one's motivations, to make a necessary contribution to a person's going blind where the only benefit is the prevention of a different person's stomach ache. So if J's act is more seriously *pro tanto* wrong than an otherwise identical act in which the agent would not have performed an independently impermissible act had different circumstances obtained, other things being equal, then this is strong evidence that CIIP is relevant in *Indifferent Contribution (Benevolent).*

[32] This is, roughly, how Walen (2005) describes intentions.

To test this, consider now a variation of the case. Suppose that, as in *Indifferent Contribution (Benevolent)*, each G and H dumps his waste to prevent a stomach ache, and each would have dumped his waste had doing so been necessary for V's blindness. However, suppose now that the third agent—K, rather than J—also dumps his waste, but would not have done so in any circumstances in which dumping the waste individually increases the risk that V, or anyone else, suffers harm. Call this case *Conditional Contribution (Benevolent)*.

It is plausible, I believe, that K (though *not* G or H) acts permissibly. Suppose that K justifies his choice to dump the waste as follows: "I am certain that V will go blind regardless of whether I dump the waste. Additionally, I am certain that dumping the waste will not result either in V suffering a harm greater than blindness or in anyone else suffering any harm. On the other hand, if I dump the waste, then I will prevent someone from suffering a comparatively small though morally significant harm. Moreover, I choose to dump the waste only *because* I am certain that my action will not individually make things worse for anyone. So I am permitted to dump the waste in order to prevent the stomach ache." This justification strikes me as compelling. Moreover, it seems like a justification that even V herself could accept. I suggest, then, that K acts permissibly.

Note that this result does not overgeneralize. Neither G nor H can make equivalent claims to justify *their* actions, for two reasons. First, neither G nor H can claim that he chooses to dump his waste only because he is certain that his action will not individually make things worse for anyone. As the case stipulates, each would have dumped his waste even in circumstances in which his contribution is necessary for V's going blind. Second, given that K would not have dumped his waste unless he could be certain that he would not individually make anyone worse off, neither G nor H can claim that V will go blind regardless of what the agent himself does. Consider, for example, that if (counterfactually) G were to refrain from dumping his waste, then K would not dump his waste either. A situation in which G refrains from dumping his waste is one where K cannot be certain that *his* dumping would not individually cause harm to V. Therefore, K would not dump his waste, and V would not suffer any harm. Therefore, G cannot assert that V would suffer the same amount of harm regardless of what G does. The same reasoning applies to H, *mutatis mutandis*.[33]

If I am correct, then while J's act is morally impermissible, K's act is morally permissible. CIIP explains why. K would have not have dumped his waste to prevent the stomach ache in circumstances in which dumping the waste to prevent the

[33] This also shows that if even two of the agents were properly motivated, such that, for example, each would have dumped his waste only in circumstances that his act does not individually increase the risk of harm to anyone, then V would not have suffered any harm.

stomach ache individually increases the risk of harm to anyone. Therefore, K would not have dumped his waste to prevent the stomach ache in circumstances in which dumping his waste to prevent the stomach ache is independently impermissible. On the other hand, J would have dumped his waste to prevent the stomach ache in many circumstances in which dumping the waste is independently impermissible.

The respect account can provide some theoretical support to CIIP and its relevance in *Indifferent Contribution* and *Conditional Contribution*. Recall that the respect account holds that acting with certain motivations can be tantamount to disrespecting a being who, in virtue of their moral status, deserves respect. And because we ought not to disrespect beings who deserve respect, an act performed with certain motivations can make the act more seriously *pro tanto* wrong than an act performed with different motivations, other things being equal.

Following many other philosophers, I claimed that harming a person as an intended means of pursuing one's ends is tantamount to disrespecting the person.[34] To harm a person as an intended means of pursuing one's ends is to regard and treat the person as an object, available for one's use, rather than as a being with ends of her own. But harming a person as an intended means of pursuing one's end is not the only way that acting with certain motivations can be tantamount to disrespecting a person. It is also disrespectful to a person to regard and treat them as a being whose ends do not matter morally, or do not matter as much as they actually do, and thus whose ends can be ignored or discounted in pursuit of one's own ends. To regard and treat a person this way is not, or need not be, equivalent to regarding and treating the person as a being who is available for one's use. Compare: an object can be something that is useful to me, but it can also just be something that can be ignored or even destroyed as I pursue my ends, given that the object has no morally significant ends of its own. Similarly, one can regard and treat a person as an object of this latter sort. To do so is to regard and treat the person as a being whose ends, even their most important ends, can be ignored or discounted, and thus can be reliably *sacrificed* as a side-effect of pursuing one's ends, even one's lesser ends.

This account can accommodate our judgments in *Indifferent Contribution (Benevolent)* and *Conditional Contribution (Benevolent)*. Because J contributes to a harm to V as a side-effect of pursuing his end, and would have contributed for the sake of that end even had his contribution been individually necessary for the harm to V, it seems that J ignores or at least significantly discounts V's interests in service of J's own ends. J regards and treats V as if V's interests do not matter, or matter very little, which is tantamount to disrespecting V. By contrast, although K contributes to a harm to V, K would have constrained his behavior had doing so made a differen-

[34] E.g., Liao 2012, Quinn 1989, Nelkin and Rickless 2014, Tadros 2011.

ce to whether or the extent to which V is harmed or at risk of being harmed. Thus, it seems that K is motivated to consider fully V's interests in deciding what to do. Thus, K's contribution does not seem to be tantamount to disrespecting V.[35]

5.

There are three further questions that I will raise in this section. I shall only be able to address each of them briefly, and my conclusions will be correspondingly sketchy. I hope to say more in the larger work.

First, should we say that K acts even *pro tanto* wrongly? Does K *wrong* V, even if he does so permissibly? If K does not act *pro tanto* wrongly, then *Conditional Contribution (Benevolent)* would serve as a second counterexample to the Bare Contribution Principle.

I am doubtful that K acts *pro tanto* wrongly. But to address this question, it will be helpful first to introduce the second question, which is this: What is the relevance to wrongdoing of an agent's contributing to a harm or risk as a side-effect of pursuing a self-interested end as opposed to a benevolent one? How, for example, should we evaluate *Indifferent Contribution (Self-Interest*) and the corresponding *Conditional Contribution (Self-Interest)*? Recall that in *Indifferent Contribution (Self-Interest*), each A, B, and C dumps his waste to avoid paying a moderate disposal cost, and each would have dumped his waste to avoid paying the cost even in circumstances in which the contribution is individually necessary for the harm to V. By contrast, in *Conditional Contribution (Self-Interest)*, the third agent, D (rather than C) would have refrained from dumping his waste to avoid the cost in circumstances in which his contribution individually increases the risk that V, or anyone else, suffers harm.

Let's compare the acts of C and D. It seems clear that C acts wrongly. It also seems that C's act is more seriously *pro tanto* wrong than D's act. As with the difference between J's act and K's act, CIIP can help explain why. D would not have dumped his waste to avoid the disposal cost in many circumstances in which dumping the waste to avoid the cost is independently impermissible, while C would have dumped his waste to avoid the cost in those circumstances (but not vice versa). Unlike C, D takes into consideration V's interests in his choice to dump his waste, and so does not seem to regard and treat V as if his interests do not matter, or matter very little. So D's act lacks a significant wrong-making feature that C's act has.

Yet D's act seems more seriously *pro tanto* wrong than K's act (in *Conditional Contribution (Benevolent)*). I suspect that this judgment primarily concerns the fact

[35] Although I have not discussed the relevance of CIIP to ordinary individuals' performance of emitting activities, I hope that the reader can anticipate—if only in broad strokes—what I will argue. As I've said, I'm developing this account for a larger work.

that the interest for which K acts is more important, morally, than the interest for which D acts, rather than whether the act is motivated by self-interest or benevolence. Holding all else equal, if D instead dumps his waste to prevent his *own* stomach ache, then it no longer seems that D's act is more seriously *pro tanto* wrong than K's act. Similarly, if K instead dumps his waste to ensure that someone else is able to avoid a monetary cost equivalent to the disposal cost that D avoids, then it does not seem that K's act is more seriously *pro tanto* wrong than D's act.

Yet the fact that D's act (dumping the waste to avoid the disposal cost to himself) seems more seriously *pro tanto* wrong than K's act (dumping the waste to prevent the stranger's stomach ache) indicates that D's act is *pro tanto* wrong. This also seems to accord with my intuitions regarding D's liability to pay compensation. D seems liable to compensate V, at least if G and H are unwilling or unable to pay.

However, we might think that D has a duty to compensate V, but still reject the claim that D owes compensation for acting *pro tanto* wrongly, i.e., for wronging V.[36] It seems possible that D's duty to compensate V is part of a more complex duty to V, which can be fulfilled *either* by (a) avoiding contributing to the harm in the first place, or (b) contributing to the harm with appropriate motivations *and* sharing the benefits of contributing with V in some way. This picture is attractive, since fulfilling the duty to V by choosing (b) would ensure that V is better off than she would have been had D not contributed (and D is better off as well). Thinking of the duty in this way also accords with the judgment that in *Palliative Contribution*, Q does not act even *pro tanto* wrongly by pouring her poison with the intention to limit V's suffering. Q could have fulfilled her duty to V either by (a) not contributing to the harm, or (b) contributing with the appropriate motivations and benefitting V in some way. By choosing (b), Q fulfills her duty to V, and thus does not act even *pro tanto* wrongly.[37]

It is more difficult to figure out how this complex duty would apply to K, if it even does. Suppose that K's duty to V is complex and can be fulfilled either by (a) not dumping the waste, or (b) dumping the waste with appropriate motivations and compensating V. Yet perhaps because K is not the one who ultimately benefits from K's act, it should be the person whose stomach ache was prevented that should pay some compensation to V. On the other hand, because K's act was not sanctioned by the beneficiary, perhaps K should make the payment after all. Still yet, perhaps we should just reject that *either* K or the beneficiary has a duty to compensate V.

These are difficult questions that I cannot resolve here. My inclination is to say

[36] At least initially; If D fails to compensate V, then D would incur a more stringent duty (of corrective justice) to compensate V in virtue of having wronged V.

[37] On this picture, the intuition that D acts at least *pro tanto* wrongly would be explained by the fact that D has not yet compensated V.

that neither K nor the beneficiary has a duty to compensate V, and that K does not act even *pro tanto* wrongly by contributing to the harm with appropriate motivations. However, I might think differently if the benefit were fungible and thus could be more easily distributed between the beneficiary of K's act and V (e.g., if the benefit was a monetary benefit, rather than a prevented stomach ache). There is obviously much more to say here.

The final question I will briefly address is whether, and how, motivations are relevant to wrongdoing when an agent makes an unnecessary contribution to a collectively-caused harm or risk, but is culpably ignorant that her act has these effects. In short, it seems that CIIP is relevant in such cases as well. An agent who causes or contributes to a harm or risk as a side-effect of pursuing some end but is culpably ignorant that her act would have these effects would have performed the act for that end across a range of circumstances, including many in which performing the act for that end is independently impermissible. Hence, the act has a significant wrong-making feature.

We can support the application of CIIP to cases of culpable ignorance by appealing to the respect account. Respecting a person involves determining, in accordance with one's abilities and available evidence, the effects that one's acts will or may have on the person. One need not necessarily aim for certainty as to whether one's acts will cause or contribute to harm or risks of harm, but one should at least take reasonable steps to determine whether they will. If one does not take these steps and one's act does cause or contribute to a harm or risk, then one treats and regards the person who is harmed or put at risk as if their interests do not matter, or matter very little, even if one's contribution is unnecessary for the harm or risk to occur. This is tantamount to disrespecting the person who is harmed.

6.

I have drawn some conclusions about how motivations can be relevant to wrongdoing in cases featuring unnecessary contributions to collectively-caused harms and risks of harm. I have not, however, said much about how these conclusions apply to the evaluation individuals' emitting behavior. That will be the project for the larger work.

References

Barry, C. & Overland, G. (2016). *Responding to global poverty: Harm, responsibility, and agency.* Cambridge: Cambridge University Press.

Botterell, A. (2007). A defence of infringement, *Law and Philosophy*.

Broome, J. (2012). *Climate matters: Ethics in a warming world*. New York, NY: W.W. Norton.

----- (2019). Against Denialism. *The Monist* 102 (1): 110–129.

Cripps, E. (2013). *Climate change and the moral agent: Individual duties in an interdependent world*. Oxford: Oxford University Press.

Draper, K. (2015). *War and individual rights*. Oxford: Oxford University Press.

Feinberg, J. (1978), Voluntary euthanasia and the inalienable right to life. *Philosophy and Public Affairs* 7.

Hale, B. (2011). Nonrenewable resources and the inevitability of outcomes. *The Monist*, 94(3), 369–390.

Hiller, A. (2011). Climate change and individual responsibility. *The Monist,* 94(3), 349–368.

Jamieson, D. (2014). *Reason in a dark time*. Oxford: Oxford University Press.

Kamm, F.M. (2007). *Intricate ethics: Rights, responsibilities, and permissible harm*, Oxford: Oxford University Press.

Lawford-Smith, H. (2016). Difference-making and individuals' climate-related obligations. In C. Hayward, & D. Rosers (Eds.). *Climate justice in a non-ideal world.* Oxford: Oxford University Press.

Liao, S.M. (2012). Intentions and moral permissibility: The case of acting permissibly with bad intentions. *Law and Philosophy,* 31, 703–724.

McMahan, J. (2009a). *Killing in war.* Oxford: Oxford University Press.

McMahan, J. (2009b). Intention, permissibility, terrorism, and war. *Philosophical Perspectives,* 23, 345–72.

Morgan-Knapp, C., & Goodman, C. (2015). Consequentialism, climate harm and individual obligations. *Ethical Theory Moral Practice,* 18, 177–190.

Nelkin, D. K. and Rickless, S. (2014), Three cheers for double effect. *Philosophy and Phenomenological Research* 89: 125–158.

Norcross, A. (1991). Intending and foreseeing death: Potholes on the road to hell. *Southwest Philosophy Review,* 15(1), 115–123.

Oberdiek, J. (2004). Lost in moral space: On the infringing/violating distinction and its place in the theory of rights, *Law and Philosophy,* 23, 325–46.

Pinkert, F. (2015). What if I cannot make a difference (and know it). *Ethics* 125 (4):971–998.

Posner, E.A., & Weisbach, D. (2010). *Climate change justice*. Princeton, NJ: Princeton University Press.

Quinn, W. (1989). Actions, intentions, and consequences: The doctrine of double effect. *Philosophy and Public Affairs,* 18(4), 334–51.

Sandberg, J. My emissions make no difference. *Environmental Ethics* 33 (3): 229–48.

Scanlon, T.M. (2001). Thomson on self-defense. In A. Byrne & R. Stalnaker (Eds.), *Fact and value* (pp. 199-214). Cambridge, MA: MIT Press.

----- (2008). *Moral dimensions: Permissibility, meaning, blame.* Cambridge, MA: Belknap Press.

Tadros, V. (2011). *The ends of harm*. Oxford: Oxford University Press.

----- (2013). Responses. *Law and Philosophy,* 32, 241–325.

Thomson, J.J. (1991). Self-defense. *Philosophy and Public Affairs,* 20, 283-310.

----- (1999). Physician-assisted suicide: Two moral arguments. *Ethics,* 109, 497–518.

Walen, A. (2006). The Doctrine of Illicit Intentions. *Philosophy and Public Affairs* 34 (1), 39–67.

Wedgewood, R. (2011) Scanlon on Double Effect. *Philosophy and Phenomenological Research* 83 (2), 464–472.

Hilary Greaves[1] and John Cusbert[2]

Comparing Existence and Non-Existence[3]

Existence comparativism holds that it can sometimes be better or worse, for a given person, that that person exists rather than not. The dominant argument in the literature on this issue is the Metaphysical Argument, which purports to show that existence comparativism is metaphysically incoherent.

We argue that the Metaphysical Argument fails. Even if existence cannot be personally better than non-existence, the Metaphysical Argument cannot be the reason for this, since the argument proves too much. Denying its first premise means holding that personal betterness comparisons between a fixed pair of possible worlds are contingent; some recent work has taken this course. Denying the second premise means holding that A can be better for S than B even when S does not exist.

We argue for this second option. First, we argue that, contrary to something of a consensus in the literature, this is not absurd in general metaphysical terms. Second, we suggest a particular analysis of personal betterness comparisons that explains in more detail how it comes about.

We conclude that existence comparativism, whether true or false, is metaphysically coherent.

[1] Global Priorities Institute & Department of Philosophy, University of Oxford, hilary.greaves@philosophy.ox.ac.uk

[2] john.cusbert@gmail.com.

[3] For valuable discussions, we are grateful to Gustaf Arrhenius, Krister Bykvist, Tim Campbell, Toby Ord, Theron Pummer, Wlodek Rabinowicz, and to audiences where earlier versions of this work were presented at the Australian National University, Sydney University and the Institute for Futures Studies. Special thanks are due to Jeff Russell, Ted Sider and Teru Thomas.

1. Introduction and motivations

Many ethical theories require us to make comparisons of overall betterness among possible worlds. Such comparisons are plausibly constrained by which worlds are better than which *for various persons*. For example, if A and B are exactly alike except that A is better than B for Jones, this suggests that A is better than B overall. It is therefore important to determine as much as we can about the extensions of the personal betterness comparisons themselves.

This essay is concerned with whether or not it can be better (or worse), for a given person, that that person exist rather than not. *Existence comparativism* (henceforth simply 'comparativism') holds that it can. On the *full comparativist* view, if a person S exists in A but not in B, then A is better (resp. worse) for S than B iff S has positive (resp. negative) well-being in A; if S has zero well-being in A and does not exist in B, then A and B are equally good for S (Roberts 1998; Holtug 2001; Roberts 2003; Adler 2009; Fleurbaey and Voorhoeve 2015). *Full anti-comparativism* holds unless S exists both in A and in B, A and B are incomparable for S (Narveson 1967; Heyd 1988; 1992; Dasgupta 1995; Broome 1999; 2004; Buchanan et al. 2001; Bykvist 2007). Both views agree that if S exists in both A and B, and has higher well-being in A than in B, then A is better for S than B. These two positions are not jointly exhaustive, but they are the most popular positions in this debate. (We discuss some additional possibilities below.)

There are three reasons why the dispute between comparativism and anti-comparativism is important. The first concerns the following principle:[4]

> **Person-Affecting Principle (PAP):** If A is better than B, then there is some possible person for whom A is better than B.

PAP is intuitively compelling, and often serves as a good guide in *fixed*-population ethics. Consider, for example, the question of whether one can improve things by 'levelling down': that is, by reducing the welfare of some person who initially has higher welfare than the average in her population, without increasing the welfare of anyone else. Such levelling down might decrease inequality. Thus egalitarians may be committed to the (perhaps problematic) conclusion that levelling-down makes things better at least 'in some respect'. It seems clear, however, that levelling down cannot make things better *overall*. And PAP seems to provide the correct expla-

[4] Outside the context of population ethics, statements of the person-affecting principle usually simply say 'person' rather than 'possible person'. Once one moves to population ethics, however, interpreting the principle in the way stated here, to include all possible persons, is near-obligatory. We defend this claim in section 3. Our statements of this and other principles are to be understood schematically.

nation of why this is so: levelling-down cannot make things better, because it does not make things better *for* anyone (Parfit 1991, 17; Temkin 1993, chap. 9; 2003a; 2003b; 2012, chap. 3).[5]

As is well known, however, if anti-comparativism is true, then PAP leads to counterintuitive results in certain nonidentity cases (Parfit 1984, chap. 16). For example, let C be a world in which everyone has extremely high well-being, and let D be a world in which the same number of people have lives barely worth living; let the populations of C and D be entirely disjoint. If anti-comparativism is true, then it is not the case that C is better than D for any of the C-people (and clearly C is not better than D either for any of the D-people or for anyone else), so PAP implies that C is not better than D. But this seems absurd. Therefore, if anti-comparativism is true, the apparently compelling PAP is inconsistent with sensible verdicts in nonidentity cases.

Many have concluded from this variable-population case that PAP is false, with potentially significant ramifications even for fixed-population ethics. But if comparativism is true, then this rejection of PAP is unmotivated: according to comparativism, C is better for the C-people than D, so, we can maintain that C is better than D but still accept PAP. The truth of comparativism would thus allow us to avoid the nonidentity problem within a person-affecting framework, and would render inappropriate the designation of population axiologies like (say) totalism as 'impersonal'.

Second: if one accepts comparativism, it becomes much harder to deny the Repugnant Conclusion. As is also well known, the Mere Addition Principle and Non-Anti-Egalitarianism (together with some plausible structural conditions) entail the apparently unacceptable Repugnant Conclusion:[6]

> **Mere Addition:** If everyone who exists in A also exists and has the same well-being level in B, but B also contains a positive number of additional people all with positive well-being (and no-one else), then B is better than A.

[5] A second example of the significance of 'person-affecting' ideas in fixed-population ethics concerns the choice between utilitarianism on the one hand, and *any* non-utilitarian approach to aggregation on the other. It has been argued that only on a utilitarian account is the *degree* to which one state of affairs is better than another overall proportional to the *degree* to which the first is better than the second for people, and that any non-utilitarian account is therefore objectionably 'impersonal' (Persson 2001).

[6] This is one version of the 'Mere Addition Paradox' (Parfit 1984, chap. 19; Carlson 1998). Other versions of the argument use 'at least as good as' or 'not worse than' in place of 'better than'; these distinctions are largely unimportant for present purposes.

The structural conditions in question are that 'better than' is transitive, that the well-being scale has the Archimedean property, and a domain richness assumption.

> **Non-Anti-Egalitarianism:** If Z has the same size population as B, Z has higher average well-being than B, and there is perfect equality of well-being in Z, then Z is better than B.

> **Repugnant Conclusion:** For any world A, there exists a world Z in which no-one has a life that is more than barely worth living, such that Z is better than A.

Virtually no-one is willing to deny Non-Anti-Egalitarianism. Thus those who reject the Repugnant Conclusion almost always deny Mere Addition. But if comparativism is true, then the Mere Addition Principle follows from a standard Pareto principle:[7]

> **Pareto Principle:** If A is at least as good as B for every possible person, and there is some possible person for whom A is better than B, then A is better than B.

The Pareto principle is very hard to deny: it is one of the very few just-about-uncontroversial principles of distributive ethics. Therefore, if comparativism is true, it is hard to deny Mere Addition, and correspondingly hard to deny the Repugnant Conclusion.

A third and complementary point is that if comparativism is true, then the Repugnant Conclusion is in any case far less repugnant. If A is even slightly worse than Z for sufficiently enormous numbers of people (viz., those who in Z have lives that are worth living and in A do not exist), then person-affecting reasoning (together with a reasonable approach to aggregation over persons) seems to weigh against A and in favour of Z. This point can of course be taken either way: as a further reason to doubt comparativism, or as a further reason to doubt one's anti-repugnance intuitions (or the required aggregative assumptions). Either way, the central point for our purposes is that comparativism is an important thesis: much else of importance to population ethics may hang on it.

Intuitions regarding the truth and falsity of comparativism go both ways. Some find comparativism intuitively compelling: if Peggy has a good life, they think, then Peggy is lucky to have been born, which makes sense provided the actual world is better for Peggy than a world in which Peggy doesn't exist. To others, the intuitive position is instead that while the actual world is better for Peggy than a world in which

[7] Here and throughout, the Pareto principle we discuss is one half of what is normally called the Strong Pareto principle. We omit consideration of other Pareto principles for conciseness.

We formulate the principle in terms of goodness, as opposed to preferences; Broome (2004) calls the 'goodness analogue' of the Strong Pareto principle the 'Principle of personal good'.

As in the case of the person-affecting principle, we write 'possible person' in our statement of this principle, in place of 'person' as is normally done outside the context of population ethics.

Peggy exists with fewer goods (less pleasure, health, knowledge, and so on), a world in which Peggy doesn't exist is a different matter, since nonexistence entails not only that Peggy doesn't possess such goods, but also that she isn't there to lack them.

The Metaphysical Argument, as we will call it, purports to establish that anti-comparativism follows from basic matters of metaphysics, so that there is no work for such substantive evaluative intuitions to do here. The purpose of this essay is to argue against this position. We will argue that the Metaphysical Argument should not sway us either way on the question of existence comparativism.

The structure of the essay is as follows. Section 2 sets out the Metaphysical Argument, identifying principles that we will call *Limited Invariance* and *Ontological Commitment* as its premises. Many have taken this argument to establish anti-comparativism. We show that this line of reasoning is not cogent: the key premise of *Ontological Commitment* is in fact inconsistent with full anti-comparativism (as well as with full comparativism). Further, the argument proves too much: accepting both premises of the argument leads naturally to conclusions that are near-uncontroversially absurd.

We turn then to the possibility of denying *Ontological Commitment*. This is the route we will eventually take. But we must proceed with care: *Ontological Commitment* seems at first sight a natural expression of the idea that only actuals, not mere possibilia, have ontological status. To date, this line of thought has not received an adequate reply.

The main aim of this essay is to supply the needed reply. Before taking up this task, section 3 clarifies the coherent alternative. Since *Ontological Commitment* and an Invariance principle jointly prove too much, if one is not to deny *Ontological Commitment* then one must deny the Invariance premise. This is the route taken by *variantism*. A variantist view might be either comparativist or anti-comparativist: the basic strategy is to agree with the full comparativist (resp. the full anti-comparativist) whenever doing so would not violate *Ontological Commitment*, but to defer to the latter whenever that imposes any restriction on personal betterness comparesons. Variantism suffers from no terrible internal flaws, but is unmotivated if (as we will argue) *Ontological Commitment* is unmotivated.

The remainder of the essay explores how *Ontological Commitment* might fail, and the implications for the debate over existence comparativism. Section 4 explores several principles that might seem natural applications of the idea that 'only actuals have ontological status' to the present case, and concludes that none succeeds in establishing *Ontological Commitment*. In several of the cases, however, plausibly *Ontological Commitment* can fail only if the grammatical and semantic structure of sentences of the form '*A* is better for *S* than *B*' come apart, in ways that have particular features.

Section 5 develops in more detail an analysis according to which the required conditions are indeed met. We call this the *Lives Account.* According to the Lives Account, sentences of the form '*A* is better for *S* than *B*' more fundamentally express not a three-place relation with *S* as one relatum, but instead a two-place relation holding between two 'ways' (or 'lives'): the way things would go for *S* if *A* obtained, and the way things would go for *S* if *B* obtained. We argue that such an analysis is independently motivated, and that on the account in question, nothing in the general metaphysics presents any obstacle to full comparativism (or to full anti-comparativism). Section 6 replies, on behalf of a Lives Account, to an argument that Krister Bykvist has offered in favour of the usual practice of interpreting '*A* is better for *S* than *B*' simply as a three-place relation with *S* among its relata.

Section 7 considers a different analysis: one that reduces personal betterness *comparisons* to prior ascriptions of monadic personal goodness *amounts*. This suggests a different argument against existence comparativism, which we will call the Well-Being Argument. The Well-Being Argument also invokes a premise that is somewhat in the spirit of *Ontological Commitment*, but avoids the pitfall of thereby 'proving too much' that plagues *Ontological Commitment* itself. We argue here, though, that the Well-Being Argument begs the question against existence comparativism.

Section 8 summarises. We do not conclude that existence comparativism is true, but we do conclude that there is no reason *of general metaphysics or semantics* why it cannot be true.

2. The Metaphysical Argument against existence comparativism

2.1 The argument

The Metaphysical Argument is succinctly stated by Broome:

> [I]t cannot ever be true that it is better for a person that she lives than that she should never have lived at all. If it were better for a person that she lives than that she should never have lived at all, then if she had never lived at all, that would have been worse for her than if she had lived. But if she had never lived at all, there would have been no her for it to be worse for, so it could not have been worse for her. (Broome 1999, 168)

Here is our official formulation of the argument:[8]

> **Limited Invariance:** If it is possible that A is better for S than B, then A would be better than B both if A obtained and if B obtained.
>
> **Ontological Commitment:** Necessarily, if A is better for S than B, then S exists.

Therefore,

> **Betterness Requires Double Existence:** If it is possible that A is better for S than B, then S exists both in A and in B.

This is not precisely Broome's formulation, but we take it that our reconstruction is faithful to the intended spirit of the argument (and that it is the argument that many other discussants have at least sometimes had in mind; we take these to include, for instance, Holtug (2001, sec. 4), Bykvist (2007), Arrhenius and Rabinowicz (2015) and Fleurbaey and Voorhoeve (2015)). A different argument, which might (with somewhat more effort) be read into Broome's paragraph, is the Well-Being Argument that we discuss in section 7 below.

2.2 Accepting anti-comparativism

The conclusion of the Metaphysical Argument — *Betterness Requires Double Existence* — implies anti-comparativism, and seems inconsistent with comparativism.[9] Many authors accept anti-comparativism, and do so apparently on the basis of the Metaphysical Argument.

A little closer inspection, however, shows that this line of reasoning is not cogent. For *Ontological Commitment*, by itself, rules out both comparativism *and anti-comparativism*, as those theses are normally understood (and as we stated them above).[10] To bring out this point, consider

[8] Our formulation of the argument deals with 'better for' statements, rather than also discussing 'worse for' and 'equally as good as for' statements, only by way of abbreviation. Everything in our discussion would equally apply, *mutatis mutandis*, to sentences of these other forms.

[9] Why only 'seems'? The point is that if persons exist necessarily, then the consequent of *Betterness Requires Double Existence* is trivially true — as are both comparativism and anti-comparativism, as stated above. Cf. footnote 9.

[10] Thanks to Teru Thomas for pressing this central point.

Example 1. Connie exists both in A and in B, though not in the actual world @. She is better off in A than in B:

	@	A	B
Connie	-	20	10

In Example 1, full comparativists and full anti-comparativists alike hold that A is better for Connie than B. But Connie does not actually exist. So, it follows from *Ontological Commitment* that A is not (actually) better for Connie than B.

Although inconsistent with both full comparativism and full anti-comparativism, perhaps *this* conclusion is acceptable; section 3 discusses views that accept it. However, things are otherwise when *Ontological Commitment* is combined with the thought that the truth-values of personal betterness comparisons should not vary from one world to another.

This brings us on to our second point. The Metaphysical Argument, as we have formulated it, assumes a premise of *limited* invariance: propositions of the form "A is better for S than B" must be true at both A and B, if true anywhere.[11] But there seems no reason to believe this without also believing a stronger condition of *full* invariance, according to which the truth-values of propositions of this form are invariant across *all* possible worlds. In Example 1, all parties to the present discussion want to agree at the very least that *if A or B had obtained*, A would have been better for Connie than B. But (as we noted above) *Ontological Commitment* prevents the proposition that A is better for Connie than B from being true at any world in which Connie does not exist; if so, a principle of full invariance would then prevent it from being true at A or at B either. That is, we are driven to the conclusion that it is impossible that any world is better than any other for any contingent being. Assuming (as we will) that persons exist merely contingently, that is absurd.[12]

[11] True *at* both A and B, not *in* A and B. The notion of truth *in* a world raises additional complications that are irrelevant for our purposes; cf. footnote 16.

[12] This assumption is widespread, but not universal. According to necessitism, *everything* exists necessarily (if at all); this includes persons alongside, for instance, atoms, rocks, institutions and shapes. A necessitist will still, of course, recognise that *some* line can be drawn roughly where we normally say the line between existence and mere possibility lies. According to Williamson (2002; 2013), it is the divide between the concrete and the non-concrete. A modal realist position in the spirit of David Lewis' (1986) might also be necessitist; the divide in that case is between the actual and the non-actual, understood indexically.

It is natural to think that if necessitism is granted, then all the obstacles to full comparativism and full anti-comparativism discussed in this paper simply evaporate. (Something like this suggestion is made by Williamson himself regarding the comparativism/anti-comparativism debate (2013, sec. 1.8), We conclude that on pain of either motivational incoherence or absurdity, one must anyway reject at least one of the premises of the Metaphysical Argument, whatever one's views on existence comparativism. There is plenty of territory to explore here, in that both of the premises of the Metaphysical Argument are at least *prima facie* very natural. But the open question can only be *in precisely what way* one or both of the premises fails, and whether that more detailed account turns out to support or to undermine existence comparativism — or, as we will argue, neither.

2.3 Denying *Ontological Commitment*

Might one deny *Ontological Commitment*?

The intuitive reason for thinking not, we take it, is that *Ontological Commitment* seems to follow from the metaphysical actualist idea that only actuals, not mere possibilia, have ontological status. Mere possibilia do not exist, and so — surely? — are not there to have any properties at all, the property of *A* being better for oneself than *B* being just a special case of this.

A very few contributors to the debate over existence comparativism have resisted this line of thought. An example is Fleurbaey and Voorhoeve, who deny the following principle:

> **No Properties of the Never-Existent:** An individual who never exists cannot have any properties, not even the relational property of something being better or worse for her. (Fleurbaey and Voorhoeve 2015, 98)

This principle seems to have struck most discussants, however, as non-negotiable. And this is understandable. For, at least at first sight, denying it would seem (for example) to undermine too much of our usual account of how we can read ontological commitments off from data about which sentences are held true. The reason why the sentence 'Tom sees Mary cannot be true unless Tom exists, for instance, seems to be that this sentence is asserting that a certain two-place relation holds between Tom and Mary, and that relations cannot hold unless their relata exist.[13]

We agree with Fleurbaey and Voorhoeve on this: the principle they call 'No Properties of the Never-Existent', if interpreted so that it entails *Ontological Commit-*

though the argument that he has in mind is closer to the Well-Being Argument that we discuss in section 7.) We are tentatively skeptical of this, but we lack the space to explore that here. In any case, for the remainder of the paper, we will largely set aside the possibility of necessitism.

[13] It is incumbent on a necessitist to further explain why 'Tom sees Mary' cannot be true unless Tom is *concrete*; cf. footnote 9.

ment, is false (as is *Ontological Commitment* itself). But more needs to be said if we are to dispel the sense that these claims are absurd. The main purpose of the essay is to execute this task.

3. Variantism

Let us first clarify the coherent alternative. What we call the *variantist* response accepts *Ontological Commitment,* but resists the Metaphysical Argument by denying *Limited Invariance*.

Consider again Example 1 from the previous section. If *Ontological Commitment* holds, then '*A* is better for Connie than *B*' cannot be *actually* true: Connie does not actually exist. If we deny full invariance, however, this does not prevent the proposition in question from being true at worlds at which Connie *does* exist, including *A* and *B*. According to the variantist, this possibility is realised: *If A or B had obtained,* then *A* would have been better for Connie than *B*.

The positions we are interested in go further, and also deny the weaker principle of *Limited Invariance*. The resulting variantist positions might be either comparativist or anti-comparativist in inclination. To see the difference, consider an example of the type that takes centre stage in the dispute between comparativists and anti-comparativists:

Example 2. Peggy exists and is happy in @, but does not exist in *B*.

	@	*B*
Peggy	10	-

A variantist, of course, will insist (as per *Ontological Commitment)* that if *B* had obtained, then @ would not have been better for Peggy than *B*. But on the question of whether @ is *actually* better for Peggy than *B* , both comparativist and anti-comparativist options are open. The position that we will call *comparativist variantism* (resp. *anti-comparativist variantism*) reproduces the verdicts of full comparativism (resp. those of full anti-comparativism) whenever these are consistent with *Ontological Commitment,* and defers to *Ontological Commitment* otherwise. So, for example, the comparativist variantist holds that @ is (actually) better for Peggy than B, although it would not have been if B had obtained.[14]

We will also have occasion to consider a ‘mirror-image’ of Example 2:

> **Example 3**. Jenny exists and is happy in B, but does not exist in @.

	@	B
Jenny	-	10

On either variantist position, of course, B is not better for Jenny than @ (though the comparativist variantist holds that it would have been, if Jenny had existed).

Whether anything of importance hangs on the dispute between a variantist position and the corresponding invariantist one depends on precisely what are the associated principles linking personal to overall betterness (hereafter, ‘link principles’). As in the discussion in the introduction, the key link principles are the Pareto and Person-Affecting principles. Suppose first that these principles consider only actual people and actual personal betterness relations:

> **Actualist Indicative Pareto Principle**: If A is at least as good as B for every actual person, and there is some actual person for whom A is better than B, then A is better than B.

> **Actualist Indicative Person-Affecting Principle**: If A is better than B, then there is some actual person for whom A is better than B.

The actualist indicative link principles lead to unacceptable conclusions even aside from an insistence on *Ontological Commitment*. For instance, consider again Example 1, and let us take it that no-one except Connie exists in either A or B. Then there is no actual person for whom A is better than B: the only otherwise plausible candidate is Connie, and she is not actual. So the above Pareto principle fails to imply that A is better than B; the above Person-Affecting-Principle goes further, and implies that A is not better than B. While this conclusion might be rationalised by appeal to the ‘moral actualist’ idea that only the interests of actual persons matter

[14] Comparativist variantism is defended by Arrhenius and Rabinowicz (2015). Remarks in the same spirit are also made by Nagel (1970, 78) and by Holtug (2001, sec. 5). We are not aware of any literature defending (or indeed discussing) anti-comparativist variantism.

morally, they are otherwise highly counterintuitive. We will take it for the purposes of this essay that such strong moral actualism is unacceptable.[15]

If so, the actualist indicative link principles must be wrong. Note again that this line of reasoning does not make use of any premises concerning the extension of the personal betterness comparisons. In particular, therefore, it is independent of whether or not one accepts *Ontological Commitment*, and so independent of variantism.

The culprit in the above example, clearly, is the restriction to actual persons. This suggests the following possibilist link principles (as in section 1):

> **Possibilist Indicative Pareto Principle**: If A is at least as good as B for every possible person, and there is some possible person for whom A is better than B, then A is better than B.
>
> **Possibilist Indicative Person-Affecting Principle**: If A is better than B, then there is some possible person for whom A is better than B.

If (as full comparativism and full anti-comparativism agree) A is better for Connie than B, then the Possibilist Indicative Person-Affecting Principle (correctly) refrains from ruling out that A is better than B. If A and B are equally good for possible persons who exist in neither of these states of affairs, then the Possibilist Indicative Pareto Principle goes further, and (again, correctly) implies that A is better than B.[16]

If *Ontological Commitment* is accepted, however, then 'going possibilist' with one's link principles is not enough to avoid the unwanted moral-actualist conclu-

[15] Moral actualism is not to be conflated with the *metaphysical* actualism that motivates much of the discussion of the present essay. For discussion of moral actualism in the context of population ethics, see (Hare 2007; Greaves 2017, sec. 5.3; Arrhenius ms, chap. 9).

Let us anticipate a possible objection. It might be argued that the conclusion that A is not better than B in Example 1 is after all acceptable, on grounds that no moral-actualist *normative* implications need follow from it. Why think the latter? Well, since it is essential to the verdict that neither A nor B is actual, there seems little danger that the variantist position will imply that someone did no wrong in choosing B over A — *ex hypothesi*, nobody chose B (or A). However, this is not sufficient to block normative implications. An agent might make a choice that commits her to B conditional on some event E, rather than to A conditional on E. Absent some further factor to justify the choice, this choice seems wrong *even if, as things turn out, E does not occur*. So objectionable normative implications can easily follow from the claim that A is not better than B; it is indeed important to have link principles that avoid this conclusion.

[16] If instead A and B are incomparable for possible persons who exist in neither A nor B — a position that seems coherent, and somewhat in the spirit of anti-comparativism — then the Possibilist Indicative Pareto Principle as stated is silent on Example 1. We could get stronger implications by replacing the first clause of this principle with the weaker condition 'if there is no possible person for whom B is better than A', though the resulting principle might be less compelling in the presence of widespread incomparability.

sions regarding overall betterness. For, again, on any view that accepts *Ontological Commitment*, A *isn't* (actually) better for Connie than B. Rather, on a variantist view, it is only that *if A or B had obtained*, then A *would have been* better for Connie than B. So, to get the desired implications (and absences of implications) to follow from a variantist view, we need versions of the principles that are not only possibilist, but also subjunctive:

> **Possibilist Subjunctive Pareto Principle**: If for every possible person S, A could have been at least as good for S as B, and there is some possible person for whom A could have been better than B, then A is better than B.
>
> **Possibilist Subjunctive Person-Affecting Principle**: If A is better than B, then there is some possible person for whom A could have been better than B.

Even on a variantist view, this Person-Affecting Principle refrains from implying that A is not better than B, and (modulo the issue discussed in footnote 13) this Pareto principle implies that A is better than B — as desired.

To sum up: on pain of moral actualism, any population ethicist must allow link principles to quantify over merely possible persons, and the variantist must (in addition) opt for subjunctive versions of the link principles. Since the subjunctive principles are themselves reasonably elegant, and have a clear motivation in terms of the denial of invariance, however, this is perhaps no significant cost of the variantist view.[17]

To say that variantism does not suggest any (new) unacceptable conclusions regarding overall betterness, however, is not to say that variantism is correct. In fact, we will argue, the motivation for variantism — the sense, that is, that *Ontological Commitment* is unassailable, or even probable — springs from a misapplication of the relevant metaphysical ideas.

4. Metaphysical Actualism

The issues posed by Example 2 on the one hand, and Examples 1 and 3 on the other, are importantly different. In Example 2, *Ontological Commitment* requires that *if Peggy had not existed*, then @ *would not have been* better than B for Peggy — but actually, Peggy does exist. In Example 1 (resp. Example 3), *Ontological Commitment* requires that A (resp. @) *is* not better than B for Connie (resp. Jenny), on the

[17] Up to largely aesthetic differences in the statements of the link principles, the moves above are essentially the ones made by Arrhenius and Rabinowicz (2015), though they do not consider cases like Example 1, or the possibility of anti-comparativist variantism.

grounds that Connie (resp. Jenny) *actually* does not exist. Let us set out these two principles explicitly:

> **OC1:** If S does not exist, then A is not better for S than B.
>
> **OC2:** If S does not exist in A, then if A had obtained, A would not have been better than B for S.

OC2 is just a reformulation of *Ontological Commitment* itself; OC1 is a weakening. Ultimately, we will argue that both of these principles should be rejected. But the metaphysical issues raised by the strengthening from OC1 to OC2 are somewhat distinct from those that are already raised by the weaker principle OC1, so we will treat them separately.

In both cases, the rejection of the principle at least arguably presupposes the availability of a suitable reanalysis of the sentences of interest, so that at a deeper level, the truth-conditions of these sentences do not involve a relation obtaining with the person S as relatum. We will further indicate our own proposal for an (independently motivated) reanalysis that has these features; that is the task of section 5. But clarity is served by also understanding the workings of the language we normally speak, not only the 'deeper level'. The present section focusses on that task.

4.1 Property actualism

We have so far worked fairly directly with the actualist slogan that 'only actuals, not mere possibilia, have ontological status'. But there are several more precise theses that this slogan might suggest. One is

> **Property actualism**: For any object x and property P, it is not possible that x should have had P but not existed:
>
> $$\forall x \forall P \Box (Px \rightarrow Ex).$$

The restriction to properties rather than relations more generally is inessential: one might equally postulate[18]

[18] Henceforth, we will usually write of property actualism for simplicity of exposition, but everything we say will apply equally to relations actualism more generally.

We take the term 'property actualism' from Fine (1985). Plantinga (1983) calls the same view 'serious actualism'; Williamson (2013, sec. 4.1) calls it 'the being constraint'.

> **Relations actualism**: For any object x and relation R, it is not possible that x should have stood in R to anything but not existed:
>
> $$\forall x_1 \dots \forall x_n \forall R \Box (R x_1 \dots x_n \rightarrow E x_1 \wedge \dots \wedge E x_n).$$

These theses, if true, would ground both OC1 and OC2. However, as is well recognised in the literature on metaphysical actualism, *property (and relations) actualism is false,* unless the notion of 'property' is carefully restricted (see e.g. (Fine 1985)). To see why, consider the property of not existing. According to property actualism, it is not possible that Socrates should have had the property of not existing but not existed. Yet this seems just a convoluted way of saying that it is not possible that Socrates should have not existed — that is, that Socrates exists necessarily. But this seems absurd. (Furthermore, if the 'absurd' conclusion is true — if, that is, necessitism is true — then property actualism, while true, loses its bite.)

One might insist that nonexistence, and other features that do not presuppose existence in the way insisted by relations actualism, is not a genuine *property*; perhaps it is instead a mere 'condition' (Plantinga 1983). This insistence could be related to the distinction between grammatical and semantic structure. Grammatical structure is relevant for the purpose of assessing a sentence as grammatically correct or incorrect; semantic structure is relevant for determining ontological commitments; the two can come apart. One who says 'the average woman has 2.3 children' is not ontologically committing to the existence of an average woman, for all that the term in question is grammatically a noun phrase. Rather, significant reanalysis is required to identify the semantic structure of, and thence the ontological commitments of, this sentence.

The thought, then, is that the semantic structure of 'Socrates does not exist' is (something like) $\neg Es$. And the term 'property' is (on this account) to be reserved for items that occupy the corresponding position in the sentence's *semantic* (not grammatical) structure. So (on this account) this sentence is not attributing a property to Socrates; rather, it is *denying* that x has a particular property, and 'x does not exist' is a mere *condition* of x.

There is of course nothing objectionable about choosing to use one's technical language in that way. But if one makes this move, one must then not be too quick to move from the observation that some condition of x can be expressed in ordinary language to the conclusion that the condition in question is a genuine property. Here, we will contrast 'positive' (existence-requiring) and 'negative' (not existence-requiring) properties, rather than 'properties' and 'mere conditions', but nothing of substance hangs on this choice of terminology.

The point, then, is that while perhaps (for all we have said so far) it is not possible that there be any x that possesses some property but does not exist — $\Box\forall x\forall P(Px \rightarrow Ex)$— this relatively innocuous thesis is crucially distinct from property actualism. For property actualism rules out even that (actual) individuals can possess any properties *at other possible worlds* in which they don't exist, and that is what created the trouble about Socrates.[19]

Nonexistence is the canonical example of a negative property, but it is far from the only one. We noted above that 'Tom sees Mary' cannot be true unless Tom exists; *seeing Mary* is a positive property. But *not seeing Mary* is a negative property: if Tom had not existed, he would not have seen Mary. Similarly for *being such that Ameena's house is free of oneself*: if Tom had not existed, Ameena's house would have been free of Tom. One should not be convinced by an argument that 'If Tom had not existed then there would have been no him for Ameena's house to be free of, so Ameena's house could not have been free of him'.[20]

Similarly, many explicitly modal properties, e.g. the property of possibly seeing Mary, are such that individuals can possess those properties even at worlds in which the individual does not exist. If Tom hadn't existed, it would still have been *possible* that Tom sees Mary, since Tom's existence would still have been possible.[21]

So there are at least some negative properties. We must therefore countenance the possibility that 'A is better for S than B' expresses a relation that is negative with respect to S. If it is, then actualist scruples perhaps still underwrite OC1, which we have so far said nothing against (and which seems to follow from the more innocuous principle $\Box\forall x\forall P(Px \rightarrow Ex)$). But the modal contexts involved in OC2 are a different matter. Consider again, in Example 2: 'If Peggy had not existed, that would have been worse for her than the actual state of affairs'. According to OC2, this statement is false (or even 'absurd' (Arrhenius and Rabinowicz 2015)). But if we are dealing with a negative relation, it could be true. Peggy (by hypothesis) actually exists, and the statement says something about how things would have been with

[19] Let A be a world in which Socrates does not exist. A distraction in the present discussion is that the proposition that Socrates does not exist arguably itself does not exist in A (since, arguably, Socrates is a constituent of that proposition; see section 4.2). This generates one sense in which the proposition that Socrates does not exist would not have been true if A had been actual: if the proposition hadn't existed, *a fortiori* it wouldn't have been true either. But, clearly, there is also another sense in which it would have been true: the proposition actually exists, and what it says accurately describes one aspect of A. We might say that the proposition is true *at* A, though it is not true *in* A: by stipulation, truth in A, but not truth at A, requires that the proposition in question exists according to A. Our discussion concerns truth-at, not truth-in. (Fine (1985, 192) discusses the same distinction in terms of 'inner' and 'outer' truth.)

[20] Thanks to Jeff Russell for this latter example.

[21] A similar example is considered in the context of the present debate by Fleurbaey and Voorhoeve (2015, 98).

respect to her if she had not existed — just, perhaps, as ‘If Tom had not existed, Ameena’s house would have been free of Tom’ says something about how things would have been with respect to Tom if he had not existed.

We must countenance this possibility, but we cannot simply *postulate* that it is realised. Of the reasonably uncontroversial examples of negative properties that we have seen, most have contained either an explicit negation (‘Tom doesn’t see Mary’, ‘Socrates does not exist’) or explicitly modal terminology (‘Tom might see Mary’). The only example not containing either of these instead contained the locution ‘free of her’, which itself is naturally analysed in terms that involve negation. This suggests that property (and relations) actualism might yet be true *of a fundamental language* (a language, that is, in which grammatical and semantic structure coincide), and that perhaps they are false of ordinary English only insofar as grammatical and semantic structure there come apart. If so, any suggestion that some property is negative is hostage to the existence of a nontrivial reanalysis explaining how that comes to be.[22]

We will argue in section 5 that indeed such a reanalysis of ‘*A* is better for *S* than *B*’ is readily available. By way of high-level preview, our claim there will be that ‘*A* is better for *S* than *B*’, like ‘Tom possibly sees Mary’, is an implicitly modal locution; furthermore, that its semantic structure is such that *S* herself does not figure as a relatum in the truth-conditions. But first, let us consider OC1.

4.2 Singular propositions

The above way of denying property actualism does nothing to impugn OC1, and therefore does not occasion any change from the variantist’s position on statements of the form ‘*A* is better for *S* than *B*’ (with no modal prefix), *where S is a merely possible person*. Thus denying property actualism is still not enough to open the door to either full comparativism or full anti-comparativism. We still apparently cannot have, for instance, that *A* is better for Connie than *B*, in Example 1. There is not, actually, any such person as Connie to possess even a *negative* property.

This suggests the following position. In Example 2, it might (or might not) be that if *B* had obtained, @ would (still) have been better for Peggy than *B* — because Peggy is actual, and so can serve to render this modal assertion true by standing in a negative relation, even though the assertion concerns a counterfactual circumstance in which Peggy does not exist. But in Example 3, analogously to Example 1, it

[22] Williamson (2013, sec. 4.1) argues that even contingentists, and not only adherents of the necessitist view that he himself favours, should subscribe to property and relations actualism. But Williamson’s discussion concerns semantic structure, not the grammatical structure of a natural language; the point that these theses are false *of ordinary language* remains secure.

cannot be that @ is worse for Jenny than B, because there is no Jenny to stand in even a negative relation.

In a certain metaphysical frame of mind, this position might seem natural. But on reflection, it is odd. Consider again the relatively uncontroversial examples of negative properties that we considered above. One was 'x is not rich'. We understand how it could be true that *if Tom hadn't existed* then Tom would not have been rich. But if this is so, it also seems highly plausible that my merely possible sister is not rich. Similarly for (e.g.) the plight of Ameena's house. Further, given the course that our discussion of negative properties has taken, this should not seem paradoxical. For (recall) we conceded that some reanalysis is necessary if we are to explain how 'x is not rich' and 'Ameena's house is free of x' come to express negative properties, and we suggested that the key would be postulating a semantic structure according to which, at the deeper level of semantic structure, what is being expressed is not of the form 'x has property ϕ' at all. But then it seems that whatever the further details of the reanalysis, the same reanalysis is likely to explain how locutions of the form 'S is not rich' and 'Ameena's house is free of S' can be true of a merely possible person. And, if so, similarly for 'A is better for S than B'.

In section 5, we will argue (in the context of suggesting a particular reanalysis) that this scenario is indeed realised. First, though, we address a concern to the effect that it couldn't possibly be realised.

A *singular proposition* is a proposition that is 'directly about' a particular individual, in the sense of having that individual as a constituent. The proposition that Plato was wise, for example, plausibly has Plato as a constituent. This is to be contrasted with propositions such as that the teacher of Aristotle was wise. The latter is also 'about' Plato in an indirect sense, in that the description 'the teacher of Aristotle' picks out Plato; but the proposition contains Aristotle and the teacher-of relation as constituents, rather than Plato himself. Plausibly, propositions cannot exist unless their constituents all exist, just as sets cannot exist unless their members do. So the proposition that Plato was wise would not exist if Plato did not; the proposition that the teacher of Aristotle was wise, in contrast, could exist without Plato (though of course in some such situations it would not be true).

Perhaps, then, the relevant actualist thought is

> **Singularity:** Statements of the form 'A is better for S than B' express propositions that are singular with respect to S.

It follows from *Singularity,* together with the auxiliary claim that propositions ontologically presuppose their constituents, that neither 'B is better for Jenny than @'

in Example 3 nor '*A* is better for Connie than *B*' in Example 1 is true (because neither of them succeeds in expressing propositions; the intended propositions fail to exist). That is, *Singularity* seems to underwrite an insistence on OC1 (though not OC2).

If *this* is the explanation of why '*A* is better for *S* than *B*' requires the existence of *S*, however, there seems no reason not to extend the thesis to cover modal contexts, thus:

> **Extended Singularity:** Statements of the form 'If *C* had obtained, *A* would have been better for *S* than *B*' express propositions that are singular with respect to *S*.

But *Extended Singularity* leads to trouble. *Perhaps* it is acceptable for '*A* is better for Connie than *B*' (in Example 1) to fail to be true, given that (once we impose subjunctive link principles) nothing untoward follows from this at the level of overall betterness. But *Extended Singularity* further implies that even the counterfactual statements 'if *A* [or *B*] had obtained, *A* would have been better for Connie than *B*' fail to be true. Given this verdict, in the case of Example 1, for instance, it follows even from the subjunctive link principles that *A* is not better than *B*. As in section 3, this conclusion is unacceptable.[23]

What has gone wrong? Well, on reflection, it's perfectly obvious that in our discussion, 'Connie' (for instance) is functioning as an abbreviation for a description (whatever is the correct account of the semantics of ordinary names for *actual* people). Recall how we introduced 'Connie'. We outlined some non-actual possible worlds, and we introduced 'Connie' as a name for the merely possible person occupying such-and-such qualitative place in the worlds thus described. And of course, nothing in this account is specific to Connie: this *has* to be the way the apparent names are functioning, whenever we use apparent names for merely possible persons (as we often do, in particular, in the context of population ethics).

The upshot is that both *Singularity* and *Extended Singularity* are false whenever they would place any restriction on the truth-values of the statements of interest. Metaphysical-actualist restrictions on personal betterness comparisons cannot be motivated via considerations of singular propositions.[24]

[23] There is a further problem for the fan of *Extended Singularity* (compared to the discussion of the otherwise similar point in section 3). Even to get the conclusion that *if A had obtained then A* would have been better than *B*, we will need somehow to bring in Connie (else there is no *explanation* of A's counterfactual superiority over *B*). It is unclear how this can be done without similarly opening the door to the possibilitity that if *A* had obtained then *A* would have been better *for Connie* than *B*, and thereby abandoning the core of the 'singular propositions' idea.

[24] Absent some other wise teacher, of course, not only the proposition that Plato was wise but also the proposition that the teacher of Aristotle was wise would not have been true if Plato had not existed. That is, even propositions that are *indirectly* about some object often require that the object in question exist in order for the proposition to be true, as (apparently) per the general principle that if *x* does not

4.3 Metaphysical actualism revisited

If not property actualism (because that is either false or toothless), and not a view about singular propositions (because that is false of the fragments of language we are interested in), what is the correct expression of the basic metaphysical-actualist idea that 'only actuals have ontological status'?

In the hands of Kit Fine (1977; 1985), actualism is a commitment to a fundamental language in which the *quantifiers* are all actualist. In that case, it seems at first sight that one cannot say things like "Connie is a merely possible person': for, taken at face value, this statement quantifies over possible persons, and not only over actual persons. This form of actualism also underwrites the above thought that the more innocuous principle $\Box\forall x\forall P(Px \rightarrow Ex)$ is non-negotiable: if the quantifier $\forall x$ is actualist (that is, it quantifies only over actual objects, and not over mere possibilia), then the principle is indeed trivially true, simply because of the more basic truth that is then expressed by $\Box\forall xEx$.

Crucially, there is no suggestion here of rejecting talk of mere possibilia across the board. Use of non-fundamental languages is not a crime, and indeed such languages are often the most perspicuous for the task at hand. And in the present case, as Fine himself is at pains to emphasise, 'talk of possible worlds and possible individuals appears to make perfectly good sense' (1985, 177). So, if actualism is true, there had better be some partial translation from the non-fundamental possibilist language to the more fundamental actualist language, preserving the coherence of at least the more obviously innocuous possibilist assertions.

There is no guarantee, though, that *everything* that is expressible in a possibilist language will also be expressible in the actualist language. So there is in principle a question of whether locutions of the form 'A is better for S than B', with S a merely possible person, are among those that survive the translation to the actualist language. It seems clear to us that they are, though we omit consideration of the details of possible translations for reasons of space.[25]

exist then x has no properties. But this thought just returns us to the demand for reanalysis that we have already recognised, and that we will address in section 5.

[25] The most popular type of actualist reduction of possibilist discourse is 'proxy reduction', in which one finds some suitable surrogate, from among actual entities, for the otherwise problematic possibilist entities. For instance, one might represent possible worlds by propositions, and possible individuals by properties (see (Lewis 1986, sec. 3) for an overview of these 'ersatzist' approaches). A different approach is outlined by Fine (1977; 1985).

5. The Lives Account

In section 4, we considered various theses of a metaphysical-actualist flavour, and we concluded that none presents a convincing argument either for *Ontological Commitment* itself, or for its weakening OC1. But at several points in that discussion, we were forced to leave a hanging thread. In all the relatively uncontroversial examples of locutions that seem to involve negative properties, or that seem to predicate something of a mere possibilium in such a way as to result in truth, we could see how grammatical and semantic structure plausibly come apart, and (crucially) it was this coming-apart that explained the otherwise puzzling feature in question. So our suggestion that '*A* is better for *S* than *B*' could exhibit similar behaviour is hostage to the availability of a suitable explanatory reanalysis.

But such a reanalysis is near at hand, and independently motivated. Suppose, by way of warm-up, that Inaaya has never eaten beans, and in fact never will. Her father urges her to commence: beans, he says, are good for her! Quite plausibly, Inaaya's father speaks truly; but how can beans do any good for anyone who never goes near them? The appearance of a puzzle here dissolves when we recognise that the content of the father's assertion is (roughly) that *if Inaaya ate beans,* her health would improve. It is, that is, an implicitly modal assertion.

Similarly, the content of '*A* is better for *S* than *B*' is at least roughly that *if A obtained*, things would be better for *S* than they would be *if B obtained*. Developing this thought leads to a reanalysis on which it is unmysterious how '*A* is better for *S* than *B*' (and related assertions) can be true in cases in which (in any of several senses) *S* does not exist.

5.1 Comparing possible lives

At a superficial level, '*A* is better for *S* than *B*' expresses a three-place personal betterness relation holding between the states of affairs (or possible worlds) *A* and *B*, and the possible person *S*. That, of course, is what gave rise to the worries surrounding *Ontological Commitment.* But perhaps at a deeper level, what is being expressed is no relation of *S* at all.

This type of explanation is indeed suggested by the idea that what is being said is more fundamentally that *if A obtained, things would be better for S than they would be if B obtained.* Rephrasing slightly: *the way things would go for S if A obtained is personally better than the way things would go for S if B obtained.* Suppose we reify these 'ways things might go for a given possible person'. Then what is being asserted, when we say that *A* is better for *S* than *B*, is more fundamentally that a *two*-place personal betterness relation holds between two of these 'ways'. At this more fundamental level S herself is not among the relata, and thereby *Ontological Commitment* can fail. Thus the absurd implications of combining *Ontological Commitment* with *Full Invariance* are easily avoided on this account. This might make the account appealing to comparativists and anti-comparativists alike.

For convenience of terminology (only), let us refer to 'ways things might go for S' as possible *lives* that S might have had. It is natural to identify these lives with *properties*. At a maximally fine-grained level, the property corresponding to the way things would go for S if A obtained includes a complete description of the corresponding possible world (the property, that is, of being such that A obtains), as well as properties that distinguish between different individuals (for example, the property of eating mushrooms every Friday). For many purposes, this extreme degree of fine-graining will not be necessary, and we could take the lives in question to be significantly coarser-grained objects. The degree of fine- or coarse-graining will be largely irrelevant for our purposes. Our only constraint is that the account must be sufficiently fine-grained to distinguish between lives that are different in respects relevant to well-being. (Which respects those are, of course, is a matter of substantive theory of well-being: hedonists, preference-satisfaction theorists and objective list theorists will disagree here.)

Here is an example, to fix ideas. In the actual world, let us suppose, Abdullah has the properties of breaking his arm by falling out of a tree when aged 6, completing a degree in clinical psychology aged 23, smelling cow parsley for the first time aged 33, and many more besides. Like all of us, he also has the degenerate properties of being such that Comet Neowise passes within 65 million miles of Earth in July 2020, and being such that 2+2=4. Let $l^{@}_{\text{Abdullah}}$ be the compound property corresponding to the conjunction of *all* the properties that Abdullah actually has and that are relevant to the individuation of his life, whichever those are. We then identify $l^{@}_{\text{Abdullah}}$ with Abdullah's (actual) life.

The Lives Account, then, postulates that the relevant deeper structure involves

- A set L of possible lives.
- A two-place personal betterness relation $\succ(j,k)$, where $j,k \in L$.[26]

[26] Within the Lives Account, we refer to $\succ$ as a 'personal' betterness relation simply to distinguish the intended sense of betterness from *contributive* betterness, i.e. from the question of which lives are such that adding them to a given state of affairs improves the state of affairs by more. This choice of terminology should not mislead us into exaggerating the extent to which the instantiation of $\succ$ in the Lives Account requires the existence of persons; we discuss below (sections 5.2–5.3) the question of whether or not it does.

Strictly speaking, the account should postulate a two-place personal 'at least as good as' relation, $\succcurlyeq$, and define the strict betterness relation $\succ$ from $\succcurlyeq$ in the usual way. We formulate the account here directly in terms of $\succ$ for ease of exposition.

If we say 'A is better for S than B', we are comparing S's life in A (l_S^A) with S's life in B (l_S^B), and asserting that the former is personally better than the latter ($l_S^A \succ l_S^B$).[27] Thus, in the (strained) sentence 'The actual world is better for Deepti than the world that would have been actualised if she had stayed at home last night', the comparison being made is between the life Deepti actually has, and a different possible life (perhaps only slightly different) that she would have had if she had stayed at home last night. Many other forms of sentence, more natural-sounding in ordinary language, have truth-conditions that are best explicated by first finding a corresponding sentence of the above form, and then applying the above principles. For instance, the (natural) sentence 'Tonya would have been better off if she'd jumped' corresponds to the (strained) sentence 'The actual world is worse for Tonya than the one that would have obtained if she had jumped'.[28]

Like the question of how finely lives must be individuated, the *extension* of the dyadic personal betterness relation $\succ$ is of course determined by substantive first-order evaluative theory: hedonists will disagree with preference-satisfaction theorists, and so on.

On a Lives Account, *Ontological Commitment* is false. The condition $l_S^A \succ l_S^B$ requires only that the lives l_S^A and l_S^B exist; S herself is not among the relata. Since lives are properties, their existence is (we take it) uncontroversial for present purposes.[29] The account provides, meanwhile, no reason to doubt *Limited Invariance*: If life l_1 is better than life l_2, then necessarily l_1 is better than l_2. If the Lives Account is correct, variantism is unmotivated.

5.2 Existence comparativism and null lives

Variantism is unmotivated, but this leaves open both invariantist possibilities: full comparativism, and full anti-comparativism. How should we choose between those?

[27] Since we are postulating a reanalysis of the sentences of interest in terms of a quite different deeper structure, rather than offering a semantic value directly for the term 'S' as it appears in the original sentence, we do not hereby commit the Lives Account to the view that persons are lives.

Krister Bykvist and Wlodek Rabinowicz have both pressed us on the issue of whether a Lives Account can give an adequate treatment of compound sentences of the form 'Peggy decides to stay at home, even though going out would be better for her', without taking on something like this commitment. We hold out hope that it can, though we lack the space to explore this here.

[28] In both cases, ultimately we will want a more sophisticated account — for instance, to deal with unit comparisons of well-being, rather than mere level comparisons. But the required moves in the direction of greater sophistication are, as far as we are aware, orthogonal to the features of the Lives Account that are important for the purposes of this essay.

[29] That is, for present purposes we can set aside any nominalist scruples. One does not usually object that 'A is better for S than B' cannot be true on the grounds that the states of affairs A and B, being putative abstract objects, do not exist. It is incumbent on the nominalist to explain how to make do without our apparent commitment to the existence of abstract objects in general.

Within the Lives Account, making sense of existence comparativism requires postulating that the set L includes at least one *null life,* corresponding to never being born. This notion of a null life might seem obviously problematic: a 'null life', one might think, is not really a life at all (it is more like the *absence* of life). But nothing should hang on our choice of technical vocabulary. Given the theoretical role of the notion of a life in this framework, postulating null lives in fact seems unavoidable. For, again, the notion of a life is supposed to represent a way things might have gone for a given person. Clearly, not having been born is one way that things might have gone for each of us — or more than one way, if other details of the worlds in which we do not exist (besides the fact of our non-existence) are relevant to how things go for us in those worlds.

Since L includes null lives, nothing *in the structure of the general framework* prevents existence comparativism from being true. We postulated above that 'A is better for S than B' is true iff $l_S^A \succ l_S^B$. If S does not exist in A (resp. in B), l_S^A (resp. l_S^B) is a null life. *If in addition* these null lives stand in relations of personal betterness to non-null lives, it follows that sentences of the form 'A is better for S than B' can be true even when S does not exist in one of the worlds in question. The natural way of filling in the detail is then to hold that any null life is personally worse than (resp. personally better than, personally equally as good as) all non-null lives that have positive (resp. negative, zero) well-being. In that case, for example, '@ is better for Peggy than A' is true iff Peggy's life in @ is worth living — as per existence comparativism.[30]

The crucial question for evaluating existence comparativism, in this framework, is whether or not null lives stand in the $\succ$ relation to any other lives. This is a substantive evaluative question, about which ethicists might disagree. But it cannot be settled by appeal to anything like a general principle of metaphysics or semantics.

5.3 Objections

Let us dispose of some objections. First, one might worry that it makes no sense to say that a person has a null life in a world in which that person does not exist, just as one might worry that it makes no sense to say that A is better for Connie than B in a world in which Connie does not exist. 'How can a person have any properties, including that of having a null life, without existing?' But, as we pointed out in section 4, this line of thought implicitly assumes that the properties in question are positive rather than negative ones. And since the property of having a (particular)

[30] Indeed, it is natural, if existence comparativism is granted, to take this to be *definitional* of the zero level of well-being, and thereby of the notion of a life being 'worth living' or 'worth not living'. See (Arrhenius ms, chap. 2), for a survey of other possible definitions.

null life is just the property of not existing (and perhaps also being such that some further conditions obtain), it is quite clear that to have a null life is to possess a negative property. This is, in fact, the canonical example that we used to introduce the notion of negative properties in the first place.[31]

A second objection is closely related. We have identified possible lives with properties — in principle, maximally specific properties. This makes them sound very much like centred worlds, as indeed they are. But we are accustomed to thinking of a centred world as a pair (A, i), where A is a world and i is a possible person *who exists in A*. Again, if lives were constrained to be this sort of object, then there could not be any null lives.

What this observation brings out is that although the 'lives' in our account are very much *like* centred worlds, there is a crucial difference. By construction of the usual notion of centred worlds, it is necessary that a person exist in w in order for there to be any centred worlds in which she is the centre and w is the corresponding uncentred world. But it is not necessary that a possible person exist in w in order for her to have a life that includes *being such that w obtains*. Again, this freedom is required if lives are to represent ways things could have gone for persons. Therefore lives cannot be identified with centred worlds as the latter are normally understood. We must stick to the more fundamental notion in terms of properties.[32]

Thirdly, and again relatedly, one will be led astray if one thinks of the 'lives' in our framework as things that can possibly be *lived* by persons. This criterion is met for non-null lives, but plausibly, one cannot *live* a null life: the notion of living is such that one cannot live anything if one does not exist. But this shows only that not all possible lives (in our sense) are things that can be lived. There is no comparable obstacle to *having* a null life — having, that is, a certain negative property — in a world in which one does not exist.

A different possible objection is that that since, on the Lives Account, persons (as opposed to their lives) do not ultimately appear as semantic values, the account implies that when we say (for instance) 'A is better for Connie than B', we are not strictly speaking *talking about Connie*; and perhaps this seems odd. But this objection is merely an echo of the thought that 'A is better for Connie than B' must express a proposition that is singular with respect to Connie. We saw above (section

[31] Indeed, it has the unusual feature of being (furthermore) what we might call a *strictly negative* property: a possible person can possess this property *only* in worlds in which she does not exist.

[32] If the details of w can matter for how well things would go for x if w obtained, where w is a world in which x does not exist, then existence comparativists would need to revise their usual account. It could not then be the case that the zero point of well-being corresponds to *the* point at which existence is equally as personally good as nonexistence, since there would not then be any unique such point. While this would complicate the overall theory, it does not raise any problems of deep principle for the comparativist.

4.2) that this insistence is misguided, at least in the case of merely possible persons. In the case of actual persons, it would be possible to flesh out the account in such a way that Peggy herself is a constituent of the proposition being expressed (via a description that is used to pick out the life in question as the ones that *Peggy* has in the worlds in question), provided one does not mind having a dual account that treats merely possible persons differently. But in the case of merely possible persons, this is obviously impossible, and we should be used to living with that. It does not follow that we cannot 'talk about' merely possible persons, in a less strict (but still perfectly respectable) sense of that phrase.

5.4 Other applications of the Lives Account

We have proposed the Lives Account, in the first instance, as an explanation of how sentences of the form '*A* is better for *S* than *B*' might be true even if *S* does not exist. But the account is independently highly plausible, and it has other uses besides this one. We will highlight two.

Firstly, the Lives Account facilitates a more elegant treatment of interpersonal well-being comparisons. To theorise about well-being in an approach that takes personal betterness comparisons to be fundamentally matters of ternary relations with persons among their relata, one starts from a collection of individual betterness orderings of worlds, one for each possible person *taken separately* (all the betterness-for-S facts *for some fixed S*). These orderings directly concern *intra*personal well-being comparisons between states of affairs; one then has to perform an additional manoeuvre in order to 'bolt on' facts about *inter*personal well-being comparisons.[33] In contrast, in the Lives Account, one starts from a single betterness-for-the-individual ordering of all possible lives (i.e. regardless of 'who lives' which lives in which possible worlds), so that interpersonal well-being comparisons are present from the start, as soon as we have any information about well-being at all. This is more elegant.[34]

Secondly, moral theories that subscribe to a principle of *anonymity* find a special affinity with the Lives Account. According to anonymity, if two states of affairs agree on the assignment of well-being levels to persons up to a bijection from the one set of persons to another — if that is, the two states of affairs agree on *how many* persons live at each well-being level, and disagree only over *who* lives at each level — then the

[33] We here assume that there are positive interpersonal well-being comparisons. This is almost universally agreed among moral philosophers, though some economists disagree. For discussion, see e.g. (Fleurbaey and Voorhoeve 2016, sec. 3).

[34] Something like this point is made by Broome (2004, 97), who also suggests that theorising in terms of lives is more straightforward, and says that he has stated his account of interpersonal well-being comparisons in terms of persons only because the latter framework is 'more familiar'.

two states of affairs in question are equally good. This might affect how one chooses to describe states of affairs in the first place. In the literature on population ethics, for instance, it is common to take 'populations' — the objects that are meant to capture everything about possible worlds that is relevant to population axiology — to be assignments of well-being levels to possible persons. But if anonymity obtains, given the theoretical apparatus of the Lives Account, one might instead identify populations with multisets of possible lives (or non-null lives). Populations in that latter sense arguably contain less information, since they do not specify *who lives* each of the lives in question. But if anonymity holds, then that omitted information is evaluatively irrelevant, in which case considerations of elegance favour omitting it. Thus, *if anonymity holds,* it is an additional nice feature of the Lives Account that it then facilitates a more elegant setup of overall moral theory. Theories that violate anonymity, however, can still be stated: we *can* talk the language of the Lives Account and nonetheless still specify who has each of the lives in question.

6. Bykvist's argument

The account we offered in section 5 explains how it might come about that *Ontological Commitment* fails. It is an essential feature of this account that at some deeper level, the truth-conditions for '*A* is better for *S* than *B'* do not involve any relation holding with *S* as relatum. This goes against an assumption that the literature against existence comparativism usually makes tacitly. Witness, for example, Arrhenius and Rabinowicz:

> [I]f we took a person's life to be better or worse for her than nonexistence, then we would have to conclude that it would have been worse or better for her if she did not exist... Clearly, this is unacceptable. Nothing would have been worse or better for a person if she had not existed. ... *A triadic relation consisting in one state... being better for a person p than another state... cannot hold unless its three relata exist*. (Arrhenius and Rabinowicz 2015, 427–8; emphasis added)

Krister Bykvist, to our knowledge uniquely in the existing literature, explicitly states and suggests an argument for the assumption in question:[35]

[35] This is not to suggest that Bykvist himself would insist the assumption is inviolable. In the passage we quote, he is merely making explicit the presuppositions of his own discussion (whose concern is with related but different matters), a move that we applaud.

> I take the value-for relations *good for, bad for* and *neutral for* at face value, that is, as genuine relations holding between states of affairs and (actual) persons — *mutates mutandi* for the comparative counterparts *better for, worse for,* and e*qually as good as for*. This gives us a neat explanation of why the following inferences are valid:
>
> > Exercise is good for me. So, there is someone for whom exercise is good.
> >
> > Everything that is good for me is bad for you. Exercise is good for me. So, exercise is bad for you.
> >
> > Exercise is better for me than binge drinking. So, there is something that is better for me than binge drinking. (Bykvist 2007, 339–40)

We agree that a satisfactory account of 'A is better for S than B' must explain the validity of obviously valid inferences. So let us consider: can the Lives Account similarly explain the data that Bykvist adduces?

Bykvist considers three value-theoretic locutions: 'good for', 'bad for', 'better for'. Since our primary interest is in comparative statements, we will consider his argument-sketch as applied to 'better for'. There are two quite different cases to consider. First:[36]

> **(I1):** Exercise is better for Krister than binge drinking. So, there is something that is better for Krister than binge drinking.

Since the Lives Account does not take better-for statements as genuine relations holding between states of affairs and actual persons, Bykvist's 'neat explanation' of the validity of this inference is unavailable on the Lives Account. But an alternative explanation is forthcoming. Recall that we suggested identifying lives with properties (viz., compounds of whichever properties that are relevant to individuation of lives). Presumably, the property E of doing exercise and the property D of engaging in binge drinking are among the properties that are relevant. Let us write ⊨ for the entailment relation among properties. The Lives Account suggests understanding 'exercise is better for Krister than binge drinking' as (roughly): in sufficiently nearby and sufficiently mutually close possible worlds A, B, if $l^A_{\text{Krister}} \vDash E$, $l^B_{\text{Krister}} \vDash D$, $l^A_{\text{Krister}} \vDash \neg D$ and $l^B_{\text{Krister}} \vDash \neg E$, then $l^A_{\text{Krister}} \succ l^B_{\text{Krister}}$. Similarly, it suggests understanding 'there is something that is better for Krister than binge drinking' as

[36] We replace the pronoun 'me' with the proper name 'Krister' because complications arising from indexicals are orthogonal to the present discussion.

(roughly): there exists a property p such that in sufficiently nearby and sufficiently mutually close possible worlds A, B, if $l^{A}_{\text{Krister}} \vDash p$, $l^{B}_{\text{Krister}} \vDash D$, $l^{A}_{\text{Krister}} \vDash \neg D$ and $l^{B}_{\text{Krister}} \vDash \neg p$, then $l^{A}_{\text{Krister}} \succ l^{B}_{\text{Krister}}$. But the latter straightforwardly follows from the former. So the Lives Account, no less than the 'face value' reading, is able to explain the validity of (I1).

We might also consider an inference that involves existential quantification over persons, as opposed to over activities:[37]

> **(I2):** Exercise is better for Krister than binge drinking. So, there is someone for whom exercise is better than binge drinking.

If it were also a datum that *this* inference is valid, then this might be a datum that the Lives Account, as we have considered developing the latter, is unable to capture. But this is no datum (unless the 'someone' quantifies over merely possible persons). For 'exercise is better for Krister than binge drinking', unless given a special reading that distinguishes it from the general locution 'A is better for S than B', seems similar in all relevant respects to 'exercise is better for Connie than binge drinking', where Connie is again some particular, but merely possible, person. And we argued in sections 4–5 that this can be true despite the non-existence of Connie.[38] So (I2) is invalid, on the reading that understands the locutions in question as being the ones with which the Lives Account deals.

We conclude that Bykvist's argument does not impugn the Lives Account.

7. The Well-Being Argument

Let us now set the Lives Account aside. The task of the present section is to briefly consider a different possible reanalysis of sentences of the form 'A is better for Peggy than B'. The basic ideas of this analysis are consistent with, but do not presuppose, the Lives Account.

The key thought for this second reanalysis is that personal betterness *comparesons* are reducible to more fundamental ascriptions of monadic personal goodness *amounts* (that is, well-being levels). This thought suggests a quite different argument against existence comparativism. We will call this the Well-Being Argument:[39]

[37] Inferences of precisely this form are not on Bykvist's list, though they are analogous to his first example involving absolute goodness ascriptions ('Exercise is good for me. So, there is someone for whom exercise is good').

[38] We concede that 'exercise is better for Connie than binge drinking' sounds a little odd if Connie is merely possible, but that is for reasons of conversational context that do not undermine the essential point. Consider instead: 'A world with a preserved climate and abundant resources is better for Connie than a world with extreme global warming.'

[39] For the purposes of this discussion, we assume a condition of full invariance (cf. section 2). This

WB1: A is better for S than B iff S has a higher well-being level in A than in B. *(Premise)*

WB2: If A is better for S than B, then there is some well-being level that S has in A, and there is some well-being level that S has in B. *(From WB1)*

WB3: A person cannot have a well-being level (even the zero level) in a world in which she does not exist. *(Premise)*

WB4: If A is better for S than B, then S exists both in A and in B. *(From WB2 and WB3)*

WB5: Comparativism is false. *(From WB4)*

Like the Metaphysical Argument, one of the thoughts underlying the Well-Being Argument also invokes the idea that an object must exist in order to possess a property. But there is a crucial difference. The Well-Being Argument does not insist that property-bearers exist *in the world of evaluation* in order for propositions ascribing properties to them to be true: only that the bearer exists *in the world that the proposition describes*. Thus, for all this argument says, there is no obstacle to "Connie has well-being level w in A" being true even at a world in which Connie does not exist, provided she exists in A.[40] This means that unlike the Metaphysical Argument, the Well-Being Argument does not prove too much: it stops precisely where the advocate of anti-comparativism wishes to stop. This may well be the argument that many anti-comparativists intend to make.

A detailed discussion of the Well-Being Argument lies beyond the scope of this essay. We note, though, that for reasons related to our above discussion of the Metaphysical Argument, the Well-Being Argument is in danger of begging the question against existence comparativism. Again, we already know (cf. section 4) that some properties are negative: not existence-requiring. And we already know that the existence comparativist's position is that either 'has well-being level zero' is one of them (so that WB3 is false), or that WB1 is not true in full generality (with cases of non-existence constituting the exceptions). Furthermore, much more so than in the case of 'A is better for Connie than B' absent any reanalysis, denying WB3 in this case seems far from ad hoc: the notion of zero well-being quite plausibly has enough 'negativity' about it make this a reasonable move.[41] A complementary point is that it

means that the truth-values of sentences of the form 'A is better for Peggy than B' do not need to be relativised to worlds. The possibility of variation of truth-values across worlds is orthogonal to the concerns of the Well-Being Argument, and would only complicate the exposition.

[40] Of course, the proposition in question is not true *in* the world in which Connie does not exist, but that is beside our point; cf. fn. 16.

[41] Roberts (2003, 178) quite explicitly denies WB3, on the grounds that a non-existent person 'has no

is far from clear that monadic ascriptions of well-being levels are in any relevant sense 'prior' to personal betterness comparisons; if instead the order of priority is the reverse, and existence comparativism is true as a feature of how those comparisons behave, then the failure of WB1 and/or WB3 follows naturally. Nothing in this argument as outlined above, or (as far as we are aware) in the existing discussion of these ideas in the literature, hints at why these positions are untenable.

8. Summary and conclusions

Existence comparativism is a controversial and important thesis in population ethics. A common view in the population ethics literature is that existence comparativism is false, and furthermore that the reason for this is given by the Metaphysical Argument. In this essay, we have argued against the second component of this view.

Firstly, the Metaphysical Argument anyway proves too much. There is no reason to believe its premise of *Limited Invariance* that isn't equally a reason to believe the stronger thesis of *Full Invariance*. Meanwhile (modulo issues of necessary existence) *Full Invariance* and the other premise of the Metaphysical Argument (viz. *Ontological Commitment*) jointly imply the absurd conclusion that no world is ever better than any other for any person. So even those who are sympathetic to the argument's intended anti-comparativist conclusion must reject one of its premises.

Ontological Commitment might initially seem non-negotiable, and seems to have struck many contributors to the debate this way. If so, the only coherent option is to deny *Limited Invariance*. This leaves open both a 'comparativist variantist' position in the spirit of existence comparativism, and an 'anti-comparativist variantist' position in the spirit of anti-comparativism. A choice between these positions would have to be settled on grounds that have nothing to do with the Metaphysical Argument.

Ontological Commitment, however, is in fact highly dubious. It follows from property actualism, and this seems to explain its initial appeal. Property actualism does indeed initially seem appealing, but we have noted (following existing work in metaphysics) that property actualism is not true in general. *Ontological Commitment* would also follow if the property of *S* that is expressed by '*A* is better for *S* than *B'* happened to be one of those ('positive') properties for which property actualism

properties at all', and that therefore these properties 'add up to a zero level of well-being'. While there is scope to say more here, Roberts' remarks should not seem paradoxical if one is armed with a positive/negative property distinction.

Arrhenius and Rabinowicz write in reply that '[h]aving a zero degree of well-being is arguably the kind of property the instantiation of which requires the existence of property bearers' (2015, 429), but they do not supply the suggested argument.

Holtug (2001) can be read as arguing that the property of having zero well-being is a negative property, on the basis of considering various substantive theories of well-being.

does hold; but we have seen that there is also strong independent reason to doubt that. Nor does any other plausible precisification of the idea that 'only actuals have ontological status' provide any good argument for *Ontological Commitment*. So the Metaphysical Argument is unsound, and variantism is also insufficiently motivated.

What we called the Lives Account develops the thought that sentences of the form '*A* is better for *S* than *B*' express that the way things would have gone for *S* if *A* had obtained is personally better than the way things would have gone for *S* if *B* obtained. It is clear how, on this account, either full comparativism or full anti-comparativism could come about. In particular, it is clear that '*A* is better for *S* than *B*' expresses a negative property of *S*.

Finally, we considered the possibility that 'betterness for S' comparisons are reducible to more fundamental matters of monadic well-being ascriptions. This is part of the line of thought behind a different argument against existence comparativism: the Well-Being Argument. We conjecture that besides the Metaphysical Argument, something like this argument is the source of much of the existing resistance to existence comparativism. The Well-Being Argument is in one sense an improvement on the Metaphysical Argument, since it does at least stop with the anti-comparativist conclusion, without going on to prove too much. Until and unless more is said, however, the Well-Being Argument simply begs the question against existence comparativism, for some of the same reasons that we discussed in connection with the Metaphysical Argument. For instance, the argument assumes that possession of the zero level of well-being is a positive property, but again, there is significant independent reason to doubt that, and no argument has been provided in favour of the assumption.

We have not attempted here to *settle* the debate as to whether or not existence comparativism is true, or whether its important upshots otherwise hold (and the authors of this essay incline towards different views on these matters). Our point is only that existing arguments against it are inadequate.

References

Adler, Matthew. 2009. 'Future Generations: A Prioritarian View'. *George Washington Law Review* 77 (5/6): 1478–1520.

Arrhenius, Gustaf. ms. *Population Ethics: The Challenge of Future Generations*. Unpublished manuscript.

Arrhenius, Gustaf, and Wlodek Rabinowicz. 2015. 'The Value of Existence'. In *The Oxford Handbook of Value Theory*, edited by Iwao Hirose and Jonas Olson, 424–444. Oxford Handbooks in Philosophy. New York: Oxford University Press.

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

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

Buchanan, Allen, Dan W. Brock, Norman Daniels, and Daniel Wikler. 2001. *From Chance to Choice: Genetics and Justice*. Cambridge: Cambridge University Press.

Bykvist, Krister. 2007. 'The Benefits of Coming into Existence'. *Philosophical Studies* 135 (3): 335–362.

Carlson, Erik. 1998. 'Mere Addition and Two Trilemmas of Population Ethics'. *Economics and Philosophy* 14: 283–306.

Dasgupta, Partha. 1995. *An Inquiry into Well-Being and Destitution*. New York: Oxford University Press.

Fine, Kit. 1977. 'Postscript: Prior on the Construction of Possible Worlds and Instants'. In *Worlds, Times and Selves,* Kit Fine and Arthur Prior. University of Massachusetts Press. Reprinted as Chapter 4 of Fine (2005).

———. 1985. 'Plantinga on the Reduction of Possibilist Discourse'. In *Alvin Plantinga,* edited by James E. Tomberlin and Peter Van Inwagen, 145–186. Dordrecht: Reidel. Reprinted as Chapter 5 of Fine (2005). Page numbers in citations refer to the reprint.

Fine, Kit. 2005. *Modality and Tense: Philosophical Papers*. Oxford: Clarendon Press.

Fleurbaey, Marc, and Alex Voorhoeve. 2015. 'On the Social and Personal Value of Existence'. In *Weighing and Reasoning: Themes from the Philosophy of John Broome,* edited by Iwao Hirose and Andrew Reisner, 95–109. Oxford: Oxford University Press.

———. 2016. 'Priority or Equality for Possible People?' *Ethics* 126 (4): 929–54.

Greaves, Hilary. 2017. 'Population Axiology'. *Philosophy Compass* 12 (11).

Hare, Caspar. 2007. 'Voices from Another World: Must We Respect the Interests of People Who Do Not, and Will Never, Exist?' *Ethics* 117 (3): 498–523.

Heyd, David. 1988. 'Procreation and Value: Can Ethics Deal with Futurity Problems?' *Philosophia* 18 (2–3): 151–170.

———. 1992. *Genethics: Moral Issues in the Creation of People*. Berkeley, Oxford: University of California Press.

Holtug, Nils. 2001. 'On the Value of Coming into Existence'. *The Journal of Ethics* 5 (4): 361–384.

Lewis, David. 1986. *On the Plurality of Worlds*. Oxford: Blackwell.

Nagel, Thomas. 1970. 'Death'. *Noûs* 4 (1): 73–80. Reprinted as Chapter 1 of Nagel (1979).

Nagel, Thomas. 1979. *Mortal Questions*. Cambridge: Cambridge University Press.

Narveson, Jan. 1967. 'Utilitarianism and New Generations'. *Mind* 76 (301): 62–72.

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

———. 1991. *Equality or Priority?* Department of Philosophy, University of Kansas. The Lindley Lecture, University of Kansas, November 21, 1991. Reprinted as Chapter 5 of Clayton and Williams (2000).

Clayton, Matthew, and Andrew Williams. 2000. *The Ideal of Equality*. London: Macmillan; New York: St. Martin's Press.

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

Plantinga, Alvin. 1983. 'On Existentialism'. *Philosophical Studies* 44: 1–20.

Roberts, Melinda A. 1998. *Child versus Childmaker: Future Persons and Present Duties in Ethics and the Law*. Studies in Social, Political, and Legal Philosophy. Oxford: Rowman & Littlefield.

Roberts, Melinda A. 2003. 'Can It Ever Be Better Never to Have Existed At All? Person-Based Consequentialism and a New Repugnant Conclusion'. *Journal of Applied Philosophy* 20 (2): 159–185.

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

———. 2003a. 'Egalitarianism Defended'. *Ethics* 113 (4): 764–782.

———. 2003b. 'Equality, Priority or What?' *Economics & Philosophy* 19 (1): 61–87.

———. 2012. *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. New York: Oxford University Press.

Williamson, Timothy. 2002. 'Necessary Existents'. In *Logic, Thought and Language,* edited by Anthony O'Hear. Cambridge: Cambridge University Press.

———. 2013. *Modal Logic as Metaphysics*. Oxford: Oxford University Press.

M.A. Roberts[1]

Does Climate Change Put Ethics on a Collision Course with Itself?

The purpose of this paper is to outline an intuitive ethics of climate change, one that understands our *maximizing* values, according to which it makes things better to make things better for people, to be tempered by our *existential* values, according to which existence is *just different*: making things better for a person by way of bringing that person into existence doesn't, on its own, make things better. Such a reconciliation, I argue, avoids the collision course we can otherwise anticipate between population ethics on the one hand and climate ethics on the other.

The work of reconciliation is commenced by reference to what we can call the person-affecting, or person-based, intuition. It's hard to get that intuition right; we need a formulation of the intuition that avoids the many pitfalls that many earlier formulations have fallen into. The principle I propose is, however, hardly immune to objection. In this paper, I consider and reply to two such objections both of which rely on the claim that probabilities are, in at least some cases, critical to moral evaluation. My counterargument will be that those objections evaporate just as soon as we clearly recognize that the probability facts underlying the one objection are very different from the probability facts underlying the other.

[1] The College of New Jersey, robertsm@tcnj.edu.

1. Plan for this paper

The purpose of this paper is to outline an intuitive ethics of climate change. Seems easy, right? We simply list our intuitions and note how they apply to issues relating to climate change—or the environment, or resource allocation, or sustainability.

But there's an obstacle. Our intuitions on any close inspection seem at odds with themselves. They include, on the one hand, the *basic maximizing idea* that (other things equal) it makes things morally better to make things better for people. (A quick note: As I use the term *person* here, it includes many non-human animals and at the same time excludes many humans.) *And* they include, on the other, what I will call the *existential constraint*, the idea that bringing a *person into existence,* however worth having that existence might be, *doesn't* (other things equal) make things morally better.

Accordingly, as I see things, working out an intuitive ethics of climate change is a matter of both recognizing well-established *maximizing* values, including the basic maximizing idea, but understanding that those maximizing values are tempered by but not, after all, inconsistent with, our *existential* values, including the existential constraint.

I consider the task of reconciling our maximizing against our existential values a matter of practical urgency. Otherwise, we find, I think, ethics on a collision course with itself. For *population* ethics—in the main, and as the area has developed over the last couple of decades—urges us to attend, first and foremost, to a seemingly unchecked collection of *maximizing* values, values that push us into bringing, for as long as the sun shall sustain life, ever more additional people into existence. In dramatic contrast, *climate* ethics, in the main, and as that more interdisciplinary area has developed in more recent years, urges us to attend to our *existential* values, and thus pushes us to *avoid* the additional pressures on the climate and on the Earth's stock of natural resources that will predictably come with a swiftly increasing human population and thus *avoid* additional misery on behalf all those people who do now or will eventually exist.

Now, you may initially resist the idea that we have, alongside our maximizing values, certain existential values. (Many theorists do.) So in this paper I start with a case, what I'll call the *Base Case,* that I think is convincing on that point (part 2).

I'll then turn to the work of reconciliation (part 3). The principle I consider central to reconciliation is rooted in the well-known *person-affecting,* or *person-based, intuition* (*PBI*). But if the intuition is well-known it's also much maligned. My own view is that its negative press owes much to poor formulation. One goal of this paper, then, is to articulate the intuition in the form of a principle that I think we can successfully defend.

And defend it I shall. Two of the most important of objections against the intuition are actually closely related and I will consider both those objections here (part 4). The first objection is based on what I will call the *Better Chance Case*. That objection rests on the compelling claim—the *better chance claim*—that a better *chance* of existence can and often does make things morally better in ways that we really can't quibble with. My sense is that the first objection captures a main reason why so many traditional maximizing consequentialists, and so many economists, find even the *question* whether the additional existence can, on its own, make things morally better perplexing. *Of course it can,* they say! For them, it's just *obvious* that the additional existence (other things equal) makes things better—it's just *obvious* that the basic requirements of rationality entail that result. I will argue, however, that it's a mistake to think that we can't consistently say *both* that the better *chance* of existence often does, in a certain way, make things better *and* that the actual *fact* of existence has no such salutary effect.

I then turn to the second objection (also in part 4). My guess is that the deep skepticism that many contemporary, post-Parfitian, consequentialists and most population ethicists share about the person-based intuition derives from this second objection, the *nonidentity problem*. I have argued elsewhere that it's a mistake to think that the nonidentity problem challenges the person-based intuition in any serious way, a mistake that, as it happens, is rooted in a misuse of the better chance claim. In this present paper, I will put a single instance of the most challenging among the nonidentity cases side by side against the Better Chance Case, and argue that out of that comparison a distinction clearly emerges, one that supports the plausible position that, although the better chance of existence often, in a certain way, makes things better, the choice under scrutiny in the nonidentity case makes things better in *no way at all*.

Conclusions are then noted (part 5).

2. Why we need to reconcile maximizing against existential values.

2.1 The collision course between climate ethics and population ethics

How do we count the cost of climate change?[2] It seems clear that it counts *against* a given climate policy that that policy will foreseeably cause future generations to suffer as a result of the disease, dislocation and depletion of natural resources that

[2] I borrow this phrase from Broome 1992.

we anticipate will come with significant increases in global temperatures. It seems clear that there is much to be said for a substantive, effective policy that would at least substantially reduce that suffering on behalf of future people.

That not insignificant proposition represents one *aspect*, one *component*, of our maximizing values, and on that proposition, I think, we can all surely agree—climate ethicist, population ethicist, traditional consequentialist, post-Parfitian consequentialist alike.

But now assume—hypothetically; this is actually a matter for empirical science to determine—that a swiftly increasing human population is identified as a significant contributor to increases in global temperatures and that the substantive, effective policy under consideration is one that would significantly *limit* population growth across the long-term future. Let's call that the "proactive population control policy." Under such a policy, many people who may well have existed under an alternate "no action policy" will be left out of existence altogether. Those people—in a future in which the "proactive population control policy" is implemented—won't literally *suffer* as a result of their nonexistence. Still, we should concede that their nonexistence *is* a kind of *loss*. *Whatever* we call it (*loss, harm*), it's a *reduction to zero* for each such person of whatever it is that makes life so precious to the one who lives. It's a reduction to zero in—we can say—the *wellbeing level* that person might otherwise accrue under the "no action policy."[3]

Among *population* ethicists, it seems that the dominant view is that those losses, the losses of the people who would remain merely possible under the "proactive population control policy," *count against* that policy. At least a little (90%? 50%?). The dominant view—the view of at least many (or most) population ethicists—is that the very *maximizing values* that tell us that the loss sustained by a future child when that child avoidably suffers under a "no action policy" counts against *that* policy *also* dictates that the loss sustained by a *possible* future child when that child never exists at all under a "proactive population control policy" must be counted against *that* policy.

The dominant view thus insists that the losses sustained by the future people who will exist and suffer under the one policy must be weighed against the losses sustained by the future people who will never exist at all under the alternate policy. The morally correct policy, then, as between the two would be the policy that sets us on the path toward the least bad future, that is, the *best*, future. The *total utilitarian*

[3] For purposes here, I leave the term *wellbeing* undefined other than to say that it is a measure of how good a given existence is *for a given person*. It's what makes life so precious to the one who lives. I also leave *person* undefined other than to note that I understand the term to include many non-human animals and to exclude many individuals that might be considered biologically human, e.g., a human body whose heart may be beating but whose cerebral cortex has ceased to function or a human embryo or an early human fetus that has yet to experience consciousness.

principle, with or without some slight tweaking, might be put to work to determine just which future that is.

In contrast, climate ethics—and here is where I think we see our *existential* values at work—is going to have a quite different view. Climate ethics is going to declare, I think, that the kind of loss that comes with never existing at all under the "proactive population control policy" just doesn't count against *that* policy at all. If, hypothetically, the "proactive population control policy" is what maximally guards against the suffering of *existing* and *future* people, is, that is, what maximizes their wellbeing on balance, then *it doesn't matter* how many people will *never* exist under that policy. It's *that* policy that we should go with without further ado.

Thus the collision between population ethics on the one hand and climate ethics on the other. To avoid the conflict is—I am suggesting—to understand *just how* our maximizing values are tempered by our existential values; it's to reconcile the one set of values against the other.

We can't just *say* that, of course. We have to *do* it. I'll avoid hiding the ball on how that's to happen by noting that there's no reason in the world we can't distinguish not *people* as having moral status or not—that would be a mistake—but rather their *losses* as having moral significance or not. Under that distinction, the loss sustained when one never exists at all is morally very different from the loss sustained when one exists (whether now or later) and has less wellbeing than one accessibly could have had. (More on accessibility below.) In both cases, there's a *loss*. (I thus concede in this connection that one way of making things better for people is to bring them into worth-having existences (*Existence Comparability*) and that nonexistence, for those people, is a *loss*.) But it doesn't follow that those losses have the same moral significance. We can agree—our *maximizing* values operating in full force—that the one sort of loss clearly does have full moral significance and still say—now bringing our *existential* values to bear—that the other sort of loss has no moral significance at all.

And indeed the version of the person-based intuition I'll introduce later exactly supports such a *Loss Distinction Thesis*.

I'll just note that it's the work of Larry Temkin that has put the task of reconciliation on the table to begin with.[4] It's his work that is convincing on the proposition that we aren't e.g. just maximizers, that we are existentialists as well, that we share a plurality of values that at various points seem to work against each other. This paper, then, can be viewed as a sort of homework assignment from Temkin: to explore how, in one little piece of the moral map, such a reconciliation might be achieved.

[4] Temkin 2012.

2.1 The Base Case

You may initially resist the idea that we really do have, alongside our maximizing values, certain existential values. (Many theorists do.)

The *Base Case* (Graph 1) is, I think, convincing on that point.[5] Here, a person p has a pretty rotten life in a given possible future f1 under a given choice c1. Even so, life in f1 bestows *some* benefits on p; p accrues *some* gains; p's life in f1 isn't, on balance, *less* than worth living. Let's stipulate that the existing p has a wellbeing level in f1 under c1 that happens to fall at *exactly* the zero level. And now let's add that p's zero wellbeing level in f1 is one that agents could have avoided on behalf of p by simply choosing c2 in place of c1 and thus bringing about f2 in place of f1. It's that fact about f1 that worries us morally, the fact that p's loss in f1 is entirely avoidable.

Now consider a distinct person, q, who never exists at all in f1 or in f2 but has a life clearly worth living in f3. We concede that the nonexistent q has a zero wellbeing level in f1 and in f2. And we note that agents could have avoided that low wellbeing level on behalf of q by simply choosing c3 in place of c1 or c2 and thus bringing about f3 in place of f1 or f2.

Do we think that f2 is better than f1? Surely we do (and here our "other things equal" condition is clearly met; no one beyond p and q is affected in any potentially morally relevant respect at all).

But now compare f2 against f3. One might think that, as between f2 and f3, the only issue is just *who* shall avoid the loss—just *whose* wellbeing is maximized. A principle of anonymity, or perhaps fairness; a principle that reflects our refusal to put

[5] Here and throughout this paper, f1, f2 etc. represent the outcomes, or possible worlds or *futures*, agents have the ability and the resources to bring about—that is, futures that exist as alternate available, or *accessible*, futures in the particular case. The concept of *accessibility* may vary depending on our interest in undertaking the evaluation. When our interest is to describe a relation of *moral* betterness between alternate futures—a *ranking* of those futures in terms of their overall moral betterness—accessibility represents more than a mere logical, or metaphysical, possibility, relative to f1, that p accrues more wellbeing or, relative to f3, that q exists. It's, rather, a relation that reflects the resources and the ability of moral agents, working as individuals or working together, to bring about (though not necessarily to make significantly probable) one future in place of another. Beyond an additional brief discussion of accessibility in part 2.5, the term is left undefined for purposes here. The futures represented in the graph are mutually exclusive and exhaust the accessible futures in the particular case. A name—p, q, etc.—written in italics and asterisked represents a person who never exists at all, and bold a person who does or will exist, in the indicated future. The utilities in the far left column indicate when one future is better for a person than another future—when, that is, a person has more *wellbeing*, on an overall lifetime basis, in one future than another. It's assumed for purposes here that the wellbeing level of a person who never exists in a given future is zero at that future—and thus that the future in which a person has an existence worth having is better for that person than a future in which that person never exists.

Thus I assume, for purposes here, *Existence Comparability*. On that assumption, it makes sense and can be true to say that a future in which a person exists is better (or possibly worse) for that person than any future in which that person never exists at all; by implication, a future in which a person never exists can be worse (or better) for that person than a future in which that person does or will exist.

anyone's interests ahead of anyone else's (other things equal)—may seem to compel us to say that f2 and f3 are equally good. And we can't plausibly avoid application of that principle on the position that p *somehow* has a moral status that q lacks. Yes, p exists across the board ("necessarily") and q doesn't. But an array of cases convinces us that that fact *doesn't* bear on q's moral status and *doesn't* mean that it doesn't matter what happens to q. Ditto, even if we add to the case that c1 is the choice *actually* made and f1 is the future that *actually* unfolds.

Graph 1: Base Case

Here and throughout, c1, c2 etc. are choices, and f1, f2 etc. accessible outcomes, or worlds or futures, that may obtain under a given choice with a specified degree of probability (based on information available to agents just prior to choice); bold indicates the specified person does or will exist, italics with asterisk that the person never exists, in the specified future. Each graph exhausts the choices available to the agents and the futures accessible under those choices for the case. Gray, where it appears, indicates the choice in fact made and the future that in fact unfolds.

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

All those points taken together may seem to make the conclusion that f2 and f3 are equally good—and that, with that, the conclusion that agents may do as they please between c2 and c3—all but inescapable.

But I don't think we think that. I think we think that the one future f2 is better than f1 *and* better than f3—that the one choice c2 is permissible and that c1 and c3 are both wrong.

And in that evaluation—if that's our evaluation—we see our *existential* values at work. If we were pure maximizers, we'd say that f2 and f3 are equally good. But we're not. We *don't* think that leaving the existing p to suffer in f3—that p's loss in f3—is *exactly counterbalanced* by q's loss in f2. While the dimensions of the two losses are identical, we *don't* think they have the same moral significance.

The Base Case should I think leave us convinced that, alongside our maximizing values, we also have certain *existential* values.

We now turn to the work of reconciling the one set of values against the other.[6]

[6] Following Broome—though his proposal was made to show that the values of fairness and equality can be accommodated within an additive approach, and mine to show that our existential values can be accommodated within an additive approach—we can distinguish between *wellbeing* and *contributive value* (which Broome calls *personal good*). Broome 2015. Accordingly, if we decide at the end of the day

3. The person-based intuition

3.1 Parfit's formulation

To fail to reconcile our maximizing against our existential values is to put ethics on a collision course with itself. It's to leave some of the dominant, foundational principles we find in population ethics at war with a plausible, indeed compelling, ethics of climate change.

The core principle that I think helps with the task of reconciliation—that in fact makes reconciliation possible—is what is called the *person-affecting*, or *person-based, intuition* (PBI). As articulated by Parfit decades ago, PBI provides that "*what is bad*" must be "*bad for someone*"—that is, for someone who does or will exist.[7] Which is just another way of embracing (a) the *existential* idea that the loss one sustains by virtue of one's never having existed at all isn't, in itself, "bad," and while (b) quietly incorporating the *maximizing* idea that no loss at all on the part of any existing or future person means no wrong done and (c) opening the door to the further maximizing idea that still *other* sorts of losses, those sustained in futures in which one does or will exist, really *are* bad. As pure *maximizers*, we would take the position that losses, however accrued, through nonexistence or otherwise, are bad, and equally so. As maximizers whose values are tempered by our *existential* values, we'll say that the losses (not the *people*, just the *losses*) count very differently from the moral point of view: one sort of loss counts fully, the other sort not at all.

PBI seems intuitive enough. But it comes with a lot of bad press—I think at least in part because it often isn't formulated with an eye to avoiding objections that now, thanks to Temkin, Parfit, McMahan and many others, seem both obvious and pressing. (See Appendix A.) What I want to do here is go immediately to a more plausible formulation, one that I think at least arguably survives some of those objections. I'll then turn to two objections that we can't simply design around—that remain in force however carefully PBI is itself formulated.

3.2 Expansive Very Narrow PBI (EVNPBI)

Adding some detail, including some detail to what it is to be "bad" and what it is to be "bad for" someone, I suggest the following principle.

that the overall values of different futures are to be calculated by reference to an *additive* principle, then we will say that the contributive values (CVs) of p's and q's existences in, respectively, f2 and f4 are exactly the same, which are both exactly the same as the CV of q's nonexistence in f3, while the CV of p's existence in f1 is lower. It being hard to fathom that the CV of p in f2 and q in f4 is *less* than zero, we shall have to say that the CV of existence for p in f1 is actually *negative* despite the fact that p's wellbeing level in f1 is non-negative, that is, zero.

[7] Parfit 1987 p. 363. Nicely formulated by Parfit, it was also brought to light by Narveson 1973.

> *Expansive Very Narrow Person Based Intuition* (*EVNPBI*). Where a future y is accessible to a future x, x is *morally worse* than y, and a choice made at x is *wrong*, *only if* there is a person p and a future z accessible to x such that p does or will exist in x and x is worse for p *than z* (where z may, *but need not,* be identical to y).

This principle—EVNPBI—is, in certain ways, *very narrow*: it's just a necessary condition, and just a condition on wrongness and worseness, not on permissibility or betterness. At the same time, to determine whether its condition is satisfied, the principle requires an *expansive* inquiry: a look at the full array of alternate futures accessible relative to the future x, and not just at the alternate future y we are working to compare x against.

3.3 Note on accessibility

EVNPBI makes use of a concept of *accessibility*. The notion of *possibility* alone will not give us the principle that we want. Consider the Three Outcome Case:

Graph 2: Three Outcome Case (Broome's Case)

wellbeing	f1	f2	f3
+10			**Billy**
+8		**Billy**	
+0	*Billy**		

Specifically, an appeal in this context to the concept of mere *possibility* won't let us take the position that f1 and f3 are equally good. For it's surely *possible* that Billy has more wellbeing than Billy in fact has in f3. But the result that f1 and f3 are equally good is seems clearly at the heart of PBI—a result any *reasonably* complete version of PBI should be able to generate.

We thus put the concept of *accessibility* to work instead.[8] Applied to the Three Outcome Case, then, EVNPBI immediately instructs that f1 and f3 are equally good. I then would take the further position, also a person-based position but outside the scope of EVNPBI, that the accessible future f3 reveals the moral deficiency in f2 that

[8] Feldman 1986, pp. 24–25. And, before that, to provide a semantics for modal logic (we never, e.g., say that *necessarily* P is false means that P is false in literally *all* possible worlds) Garson 2013, pp. 63–67.

makes f2 worse, not just than f3, but also f1. We now have a complete, person-based account of the case.[9]

EVNPBI in hand, accessibility in two, we now turn to the two objections.

4. Two objections

4.1 The Better Chance Case

According to EVNPBI, the worth-having additional existence does not, other things equal, make a given future better. It seems, however, undeniable that in certain cases the better *chance* of existence *does* make things better.

Consider, for example, the Better Chance Case. Here, Harry's wellbeing at +8 in f1 under c1 is far from maximized; f3 under c2 is far better for Harry than f1 under c1. Yet we surely can agree that c1 is permissible—that the probabilities at play in this case, the fact that the fertility pill substantially improves Harry's chances at existence, convert what would otherwise be a wrong choice into a perfectly permissible choice.

Graph 3: Better Chance Case

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

[9] For purposes here, I leave *accessibility* undefined other than to note the following points. Accessibility is a relation between possible futures, such that if one future is *accessible* relative to another future in a given case, then *necessarily* the one future is accessible relative to that future (facts, that is, regarding accessibility being built into the details of each future, or world, and the details of each world themselves holding as a matter of necessity). (We can call that principle the *accessibility axiom.*) If agents, relative to one future x, working as individuals or collectively, have the resources and ability to bring about a better future y, and we accordingly say that y is *accessible* to x, and if, in another case, agents in x *don't* have the resources and ability to bring about the better future y, and thus that y *isn't,* in that case, accessible to x, we have in effect equivocated on "x": the "x" in which the agents have the requisite resources and ability isn't identical to the "x" in which they don't.

Thus it should be clear that EVNPBI isn't in violation of the principle of the *independence of irrelevant alternatives,* or *independence* for short, according to which one future's being worse than a second future does not depend on the existence or availability of any third future. If y is accessible relative to x and *z is not,* and if its determined that x *isn't* worse than y, then, according to the accessibility axiom, there will be *no case* in which z *is* accessible relative to x and hence no violation of independence.

But if Harry's better *chance* under c1 at f1 makes the otherwise wrong choice of c1 at f1 permissible, and if we think that c1's permissibility in this case turns on chance (in combination with wellbeing) making the otherwise *worse*-for-Harry f1 *better*-for-Harry, then doesn't that better chance, itself necessarily *embedded,* alongside c1, in f1, make f1 better for Harry than, say, f3? Or f4? If we say yes, we'll then lose the implication from EVNPBI that f3 and f4 are at least as good as f1 is—and with that the implication that c2 at f3 and c2 at f4 are both permissible.

Reply to the better chance objection. I think this argument against EVNPBI works, if, but only if, we understand the probabilities critical to moral analysis to bear on our evaluation of when one accessible *future* is better for a person than another and—from there—to bear on the evaluation of *choice.* But I think the argument fails if we instead understand the probabilities to bear exclusively on the evaluation of *choice.* (See Appendix B; distinction between the affect-the-future view and the affect-just-the-choice view; to preserve EVNPBI, we adopt the latter.)

Under that latter understanding, we can say that, in many cases and in certain ways, the better *chance* of existence can make things better. Thus in the Better Chance Case the better chance converts the otherwise wrong c1 into a perfectly permissible c1. But we can also insist that the actual *fact* of existence doesn't have any salutary effect at all. Thus it *doesn't* make f1 better than f2 or better than f4. That understanding would clarify that the ranking that EVNPBI generates *before* probabilities are taken into account and according to which f2, f3 and f4 are equally good and at least as good as f1 is exactly the same as the ranking that EVNPBI generates *after* probabilities are taken into account. Moreover, EVNPBI on that latter understanding can as before leave the door open to still other person-based principles, perhaps a person-based Pareto principle, that could step in to say that, notwithstanding Harry's better chance, f1 is actually *worse* than f3 and thus worse than f2 and f4 as well.

The upshot for permissibility of that latter understanding? EVNPBI tells us *nothing* about the permissibility of c1 at f1. But it does at least generate some results for us that nicely reflect both at least some of our maximizing values and at least some of our existential values: that c2 at f3 and c2 at f4 are both perfectly permissible.

Still, it *is* a problem that EVNPBI doesn't instruct that c1 at f1 is *permissible.* It means that EVNPBI arguably isn't even a *reasonably* complete person-based principle. For a *reasonably* complete principle would surely *support,* i.e. *imply,* that c1 at f1 is permissible. To remain true to the person-based intuition, however, any worthwhile extension of EVNPBI that generates the result that c1 at f1 is permissible must also take care not to simply toss the person-based results we've just noted EVNPBI so nicely generates: that c2 at f3 and c2 at f4 are permissible as well.

To that end I relied in some of my earlier work on a person-based appeal to the concept of *expected value* (*EV*). I now think that approach is problematic. We need another concept, another vehicle for balancing the better *chance* against the better *outcome*.[10] One option is the concept of *probable value*.

First, a definition. Where a choice c made at a future f creates a probability n that p will have the wellbeing level (WB) that f in fact assigns to p, we can say that *f's probable value (PV) for p under c* is n(WB).

And now the principle:

Expansive Very Narrow Person Based Intuition + Probable Value (EVNPBI+PV):

> Where y is accessible to x, x is worse than y *only if* there is a person p and a future z accessible to x' such that
>
> p does or will exist in x and
> x is worse for p than z (where z may, but need not, be identical to y); and
>
> c at x is wrong *only if* there is a person p and an alternate choice c′ at a future y accessible to x such that
>
> p does or will exist in x and
> x is worse for p than y and
> PV of c at x for p < PV of c′ at y for p.

Applied to the Better Chance Case, EVNPBI+PV tells us that c1 and c2, wherever performed, are permissible. c1 is permissible in virtue of the fact that, wherever performed, PV(c1) is at least as great as PV(c2). And c2 performed at f4 is permissible in virtue of the fact that Harry's wellbeing in f3 is maximized. Finally, c2 performed at f4 is permissible in virtue of the fact that Harry never exists at all in f4.

That account of the Better Chance Case itself seems highly plausible, nicely reflecting our existential values while recognizing a clear sense in which a better chance at

[10] It's true that, by simply adding to EVNPBI a clause that deems a choice permissible in any case in which no alternate choice comes with more expected value than the one choice does for any existing or future person, we can easily provide a plausible account of the Better Chance Case itself. Since EV(c1) is at least as great—indeed, it's greater—than EV(c2), the new principle would immediately instruct that c1 is permissible. Since we would just add the new clause, not substitute it for the clause that lets the agents off the hook when they (at the end of the day) succeed in maximizing wellbeing for the person or when the person never ends up existing at all, the principle would also instruct that c2 (wherever performed) is permissible.

But I now think that that approach is problematic, incorrectly implying in certain cases that a given choice is permissible in virtue of a microscopic chance of a truly wonderful outcome. Where both the future that *in fact obtains* is very bad for person—perhaps the wellbeing level is even in the negative range—and the *probability* that the choice under scrutiny would produce just that outcome is very high, that choice may well in fact be wrong.

Outlier cases support this critique. This issue was pointed out to me by Dean Spears.

existence can make things better: neither the better chance of existence, nor the actual fact of existence, will make the *future* better, but it can easily make the otherwise wrong *choice* perfectly permissible.[11]

4.2 The non-identity problem

One quite challenging version of the *nonidentity problem* arises very naturally in the context of climate change.[12]

The concern is that PBI—whether understood by reference to EVNPBI+PV or any other construction of PBI that credibly takes probability into account—generates the highly implausible result that we, in effect, need do little and perhaps nothing at all to control the flow of carbon into the atmosphere since, whatever we do, and *provided* that we take *both* the chance of a given person's making it into existence as well as the wellbeing that person has in those futures in which he or she does exist, our choice will be maximizing for each *future* person who will ever exist at all. Looked at in that way, the "no carbon limit policy," the policy we anticipate will cause future people to suffer, which suffering, we said, should be counted *against* that policy, actually turns out to be *maximizing* for each such future person. For well known reasons: what would their chances of existence have been, had the future not unfolded pretty much just as it did under that "no carbon action" policy? The usual (if not clearly correct) answer: but for the "no carbon limit policy," they'd have had almost no chance of ever existing at all.[13]

[11] As it stands, EVNPBI+PV won't account for many other cases we are surely interested in. Specifically, it will remain silent in cases in which a choice is clearly permissible given the probabilities, but where PV is relatively low in virtue of the fact that the probability that any one of *two or more high wellbeing futures* will unfold is relatively low.

That EVNPBI+PV is limited—that it doesn't generate results in the sort of case—was pointed out to me by Christopher J.G. Meacham.

That we will want, however, to *extend* EVNPBI+PV does not mean that there is anything amiss with EVNPBI+PV as it stands.

[12] See Broome 1992.

[13] Thus virtually all population ethicists—indeed, virtually all contemporary consequentialists—owe their thinking in connection with that question to Parfit. On that view, the sweeping policy changes agents would have needed to put into effect, many years prior to Gloria's coming into existence, to have alleviate the suffering that climate change will otherwise impose on her would have had the effect, not of making things better for *Gloria*, but rather only of bringing *another* person—a person *nonidentical* to Gloria—into existence in place of Gloria. Yes, that person would have been better off—climate-change-related suffering, on that distinct person's behalf, having been alleviated—than Gloria is. But for Gloria herself we can only anticipate a very low probability of any better result—and a very high probability of her never existing at all.

The thinking behind that conclusion is itself rooted in what Kavka called the "precariousness" of existence (Kavka 1981, p. 93.): any little change in the history that leads up to any particular person coming into existence is likely to affect the "timing and manner" of all the conceptions that will then take place. Parfit 1987, pp. 351–379.

On that way of looking at things, EVNPBI+PV immediately generates results that seem to us to be clearly false: that the “no carbon limit policy” is perfectly permissible.

That way of looking at the case is summed up in *Climate Change/Nonidentity Case Version I* (Graph 4). Let’s suppose that agents in fact choose c1, a “no carbon limit policy” on climate change, and that the future f1 in fact unfolds. Let’s suppose, further, that Gloria is a person who does not yet but eventually will exist and suffer in f1 under c1 as a result of the disease, dislocation and natural resource shortage that take place under that “no carbon limit policy.”

Graph 4: Climate Change/Nonidentity Case Version I

	c1: choice to continue carbon emissions at current levels over next five years (“no carbon limit policy”)		c2: choice to reduce significantly carbon emissions over next five years (“proactive carbon limit policy”)	
probability	0.1	0.9	0.0001	0.9999
wellbeing	f1	f2	f3	f4
+10			**Gloria**	
+2	**Gloria**			
+0		*Gloria**		*Gloria**

It’s the *probabilities* that the case takes for granted—the probabilities displayed in Graph 4—that explain why it is thought that c1 is *maximizing* for Gloria, *despite* the suffering she endures in f1 under c1 and *despite* the fact that she is worse off in f1 than she is in f3. In the Better Chance Case, Harry’s better chance of coming into existence under c1 makes up for his lower wellbeing level under c1 at f1. Similarly here, or so the argument goes, Gloria’s better chance of coming into existence under c1 makes up for her lower wellbeing level under c1 at f1.

Reply to nonidentity objection. The comparison between the Better Chance Case and the Climate Change Case can help us, I think, identify a flaw in the thinking behind the logic of the nonidentity problem.[14] Put most simply, the “no carbon limit

As the discussion in the text below should make clear, in my view the line of argument that concludes that the choice under scrutiny, taking wellbeing and probability into account, is better for Gloria is in fact deeply flawed.

[14] I have described that flaw in the thinking behind the nonidentity problem more detail elsewhere, in connection with Parfit’s depletion and risky policy cases, Kavka’s slave child and pleasure pill cases and cases involving historical injustices. See Roberts 2007 and 2009.

policy" on climate change just isn't a fertility pill. Even as we recognize what Kavka called the *precariousness of existence,* we should be able to see that the "no carbon limit policy"— the choice of c1—*doesn't* actually increase Gloria's chances of coming into existence. That she *does,* ex hypothesi, eventually exist doesn't increase the *probability,* calculated as of that time just prior to choice and based on information then available to agents, of her existence.

The confusion has, I think, arisen because theorists naturally focus on what *does* happen, an *irrelevant* if concrete fact whether we think probabilities should be brought to bear via standard expected value theory or via my concept of probable value, and thus fail to focus on the *relevant* if abstract fact: the many ways the future *might* have unfolded under what we all agree to be the worse option—that is, the "no carbon limit policy." The fact is that, for each of the huge umbrellas of "no carbon limit policy" and "proactive carbon limit policy, implementation can be achieved in a vast number of distinct particular ways. Some of those ways will result in (perhaps they will actually *determine*) Gloria's coming into existence; many, many will result in her never existing at all. But we have no ground for thinking that the probability that Gloria will exist under c1 is greater than the probability that she will exist under c2.

If anything, the hazards the "no carbon limit action" policy on climate change will unleash on future people may well make it *less* likely, not *more* likely, that Gloria will exist than would the more secure environment that we can anticipate under the "proactive carbon limit policy."[15]

Consider, then, the following Graph 5, which provides a more accurate picture of the probabilities in fact at stake in connection with the climate change.

The critical point is that the choice of c1 leaves Gloria's chances of ever existing at all just as microscopically small (indeed, .0001 wildly *overestimates* those chances) as c2 does. For both choices, how the choice is implemented *may* determine one particular future f1 rather than another. But until the future in fact unfolds, agents

[15] Similar, the Holocaust surely made the children, grandchildren and great-grandchildren *less* likely to come into existence rather than *more* likely to come into existence. (I owe this suggestion, an extension of my point that those children were no less likely to come into existence had the Holocaust never occurred, to Peter Singer.) If a person's coming into existence is highly precariousness in any case, surely it is made more so when a powerful regime becomes intent on annihilating that person's forebears.

Clearer thinking about the probabilities at stake in the different versions of the nonidentity problem can also help us sort through the different versions of the *repugnant conclusion*. Parfit 1987, pp. 381–390. Thus, where a policy of overpopulation has strained resources, leaving vast numbers of people with lives only barely worth living, when an alternate population policy would have left fewer people with lives well worth living, the question should not be: what is the probability that all those vast numbers of people would have existed and been better off? but rather: for any one such person who has been left with a life only barely worth living, would an alternate policy, one that resulted in ample resources, have made any less likely that that person would ever have existed at all?

can’t have any clear idea just what form implementation will take. Gloria’s chances of existence are thus very, very small under c1—*just as they are under c2*.

Graph 5: Climate Change/Nonidentity Case Version II

	c1: choice to continue carbon emissions at current levels over next five years (“no carbon limit policy”)				c2: choice significantly to reduce carbon emissions over next five years (“proactive carbon limit policy”)			
probability	0.0001	0.0001	0.0001	…	0.0001	0.0001	0.0001	…
how choice implemented	i_1(c1)	i_2(c1)	i_3(c1)	…	i_1(c2)	i_2(c2)	i_3(c2)	…
wellbeing	f1	f2′	f2′′	…	f3	f4′	f4′′	…
+10					**Gloria**			
+8	**Gloria**							
+0		_Gloria*_	_Gloria*_	_Gloria*_		_Gloria*_	_Gloria*_	_Gloria*_

That means that EVNPBI+PV easily avoids the result that the choice of the “no carbon limit policy” c1 at f1 is permissible and thus opens the door for the result that that choice is wrong. For *all* of the necessary conditions on that choice being wrong—that Gloria, first and foremost, exists at f1; that an alternate future f3 create more wellbeing (more *actual* value) for Gloria than f1 does; and that that same alternate future creates more *probable* value for Gloria than f1 does—are fully satisfied.

5. Conclusion

It seems that, if we want it, we can retain the person-based intuition, in, for example, the form of EVNPBI+PV. Recognizing that principle helps us see how our maximizing values are not inconsistent with but rather are tempered by our existential values. That reconciliation, in turn, provides the basis for an intuitive ethics of climate change—one that avoids putting ethics on a collision course with itself and one that won’t obligate, or allow, us to increase the suffering of the people who do or will exist on the altar of bringing ever more people into existence but that may well require us to understand climate change as among the most pressing moral issues of our time, with critical steps to be taken, now or never, to alleviate the suffering that future people will otherwise be forced to endure.

Appendix A

> *Very Narrow PBI.* Where a future y is accessible to x, x is *morally worse* than y, and a choice made at x is *wrong, only if* there is a person p such that p does or will exist in x and x makes things *worse for* p.

One feature that makes this principle narrow is that it asserts only a condition on when a future is *worse,* not on when a future is *better*. It can only tell us that a future in which the *worse* off person never exists isn't *worse;* it can't tell us that a future in which the *better* off person never exists isn't *better*.

To see why it's important to narrow PBI in that way, consider Graph A1 below (*Wrongful Life*).

Graph A1: Wrongful Life

wellbeing	f1	f2
+0		*Payal**
-10	**Payal**	

Payal's life in f1 is thoroughly miserable; her wellbeing level is in the negative range. It's plausible to say that it would have been better for Payal herself had Payal never have existed at all and, for that reason, that f1 is *worse* than f2—that is, that f2 is *better* than f1 But if a person for whom things are better must *exist* in a given future for that future to be better, then we can't say that. Thus: the narrow construction of PBI will in the end make for a more credible principle.

A second feature further narrows the principle. The principle provides only a *necessary,* and not a *sufficient,* condition on when a future is worse. Even if we carefully limit the application of such a sufficient condition, we run into trouble in cases in which the wellbeing or existence of still other people is at stake.

To see that point, consider Graph A2 (*Double Wrongful Life*).

Graph A2: Double Wrongful Life

wellbeing	f1	f2
+0	*George**	*Payal**
-10	**Payal**	**George**

In this case, it is highly plausible that f1 is exactly as good as f2. That means that the fact that Payal exists and suffers in f1, and that her suffering can be avoided *at no cost to anyone else who does or will exist in f1*, is not *sufficient* to show that f1 is worse than f2.[16] A plausible, more complete person-based view will include person-based principles that provide sufficient as well as additional necessary conditions on worseness. But the plausible sufficient condition won't be one that mechanically tracks the mere necessary condition we have in Very Narrow PBI.[17]

A well-formulated PBI will thus be *narrow*—indeed, *very* narrow. But Very Narrow PBI itself is not a principle we can evaluate. For it's elliptical in a certain respect. A future that is *worse* must make things *worse for* an existing or future person. But worse for that person *than what*?

It may seem perfectly obvious how the ellipsis is to be completed (and has indeed seemed so to many philosophers[18]). To determine whether the condition on x being worse than y is satisfied, it may seem that PBI requires us to look no further than whether x is worse for an person who does or will exist in x *than y*. On that construction, our survey of the futures accessible to x for purposes of determining whether the condition is satisfied is *highly restricted*. Thus:

> *Highly Restricted Very Narrow PBI* (*HRVNPBI*). Where a future y is accessible to a future x, x is *morally worse* than y, and a choice made at x is *wrong*, *only if* there is a person p such that p does or will exist in x and x is worse for p *than y*.

That highly restricted construction of PBI—*HRVNPBI* for short—however, dooms PBI to failure.

Consider Graph A3 (*Three Outcome Case*).[19]

Graph A3: Three Outcome Case

wellbeing	f1	f2	f3
+10			**Billy**
+8		**Billy**	
+0	*Billy**		

[16] Other cases as well make this point. See Hare 2007. See also Roberts 2011a and 2011b.

[17] See, however, Parfit 2018.

[18] See e.g. Parfit 2017.

[19] This case is from Broome, as is the line of reasoning itself. Broome 2004, pp. 140–149. But I use that line of reasoning here to support a conclusion other than the one Broome suggests.

For PBI to remain credible, it must be understood as consistent with a highly plausible Pareto principle, one that is limited to the case where the population in one future is identical to the population of another future (we'll call it *Same-People Pareto*). According to that principle, since the populations of f2 and f3 are identical, and since f2 is *worse for* one person, Billy, and better for no one, f2 is *worse* than f3. At the same time, it's clear that f2 isn't *worse* for Billy than f1. (Since Billy has a life that is worth living in f2 and no life at all in f1, if anything, f2 is *better* for Billy than f1.) Ditto for f3 as compared against f1. HRVNPBI thus implies both that f2 isn't worse than f1 and that f3 isn't worse than f1. The principle, in addition, implies both that f1 isn't worse than f2 and that f1 isn't worse than f3. On those grounds we then infer that f1 and f2, and f1 and f3, are exactly as good as each other and, finally, that f2 is exactly as good as f3. But now we have an inconsistency. f2 can't be both *exactly as good as* f3 *and worse than* f3. Something shall have to go.[20]

But it doesn't follow that what has to go is PBI. We can consistently instead reject HRVNPBI in favor of the position that, when it comes to the existence of an additional person, and our aim is to compare one future against another, it's not enough to ask whether the person is worse off in the one future than in the other. Rather, we must ask whether the person is *avoidably* worse off in the one future. And that requires a more searching inquiry—an inquiry into whether there's *any* further alternate accessible future that is better for that person. If there is, then the condition itself is satisfied.

Thus:

> *Expansive Very Narrow PBI* (*EVNPBI*). Where a future y is accessible to a future x, x is *morally worse* than y, and a choice made at x is *wrong, only if* there is a person p and a future z accessible x such that p does or will exist in x and x is worse for p *than z* (where z may, *but need not,* be identical to y).

Since Expansive Very Narrow PBI—*EVNPBI* for short—requires an inquiry into not just how Billy fares in f2 as compared against *f1* but also how Billy fares in f2 as compared against *f3,* and since f2 is worse for Billy than f3, the necessary condition on f2's being worse than *f1* is satisfied. We thus avoid the result that f2 is at least as good as f1 and thus avoid the inconsistency. That, in turn, leaves us free to adopt still other person-based principles—here, it's enough to adopt a combination of same-people Pareto and various conceptual principles—that tell us that f2 is indeed worse than f1. EVNPBI tells us, moreover, that f1 is exactly as good as f3.

[20] Broome himself insists that it is the neutrality intuition, his own take on PBI, that we must reject. Broome 2004, pp. 140–149. I agree with him that we must reject the neutrality intuition, which instructs that f2 is *worse* than f1. But it doesn't follow that we must reject EVNPBI.

We are thus left with an account of the case that is both plausible and consistent: f1 is exactly as good as f3, while f2 is worse than both. Agents are permitted to bring Billy into existence and permitted not to bring Billy into existence, but it would be wrong to bring Billy into existence and make things worse for him rather than better. One feature that makes this principle narrow is that it asserts only a condition on when a future is *worse*, not on when a future is *better*. It can only tell us that a future in which the *worse* off person never exists isn't *worse*; it can't tell us that a future in which the *better* off person never exists isn't *better*.

Appendix B

But what do those shifting values actually work on? Does the probability—as a prior matter; foundationally—affect the value of the *future* in which the choice is made? Does it make one *future* better than another, such that a choice that ends in the one future is itself deemed better? Let's call the view that says that it does the *affect-the-future view*.

Or does it leave the valuation of the *future* alone and instead act just on the choice itself? Does the probability affect just the value of the *choice*? Let's call the view that says that that is what is going on the *affect-just-the-choice view*.

If we aren't clear on which view is correct, then we will be unable to rank one future against another in cases in which the probability that a given choice will give rise to a given future is less than 1.

Consider, again, the Better Chance Case (Graph 3).

Graph 3: Better Chance Case

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

The shaded area here and in what follows means that agents choose c1 and that c1 in fact ends in f1—in, that is, Harry's coming into existence at a wellbeing level of +8. Let's assume that c1 at f1 is permissible. That seems undeniable. While Harry's wellbeing level in f1 accessibly could have been higher—that is, it is higher f3; f3 is better for Harry than f1—the fact that c1 increases his chances of ever coming into existence at all seems clearly to counterbalance that fact and renders c1 permissible.

In addition, we should take for granted that the fact that Harry's probability of existence under c1 is 0.1 is a fact embedded in the future f1 along with c1 itself. Similarly, the fact that Harry's probability of existence under c2 is 0.0001 is a fact embedded in the future f3 along with c2 itself. Those are simply some of the many empirical details that are included in f1 and f3, respectively.

And now our question: Do those facts about probabilities increase the value of the *future* f1 itself as compared against the *future* f3? Is, that is, the *affect-the-future* view correct?

On the affect-the-future view, while Harry has less wellbeing in f1 than in f3, we still have ample room to say that, given Harry's higher probability of existence under c1, f1 is better than f3—that is, that f3 is worse than f1. Taking both wellbeing *and* probability into account, in other words, we have ample room to say that f1 is *all things considered* better than f3—that is, that f3 is *all things considered* worse than f1. Indeed, the difference in probability under c1 and c2 being so great, and the difference in the wellbeing levels themselves being relatively small, it seems that any affect-the-future principle that takes both into account will generate the result that f1 is better than f3—that is, that f3 is worse than f1.

We then compare f2 and f4. Harry doesn't exist in either f2 or f4, and so we say that Harry has exactly as much wellbeing in f2 as Harry has in f4. As noted earlier, however, the choice c1—along with the higher probability of Harry's existence that comes with that choice—is embedded among the factual details of f2. Similarly, the choice c2—along with its lower probability of Harry's existence—is embedded in f4. Under the affect-the-future view, we accordingly say that f2 is better than f4—that is, that f4 is worse than f2.

It would seem that, on that view, it's better to make the choice that increases the chances of existence for more people (there being nothing special about Harry) even if (within limits; the life must be worth living) that choice also reduces (perhaps dramatically) the individual wellbeing levels for all of those who do or will exist. Those results provide support for the surely correct view that c1 is permissible.

Moreover, since, on the affect-the-future view, the future that c1 will produce is better, whether that future happens to be f1 or f2, than any alternate future c2 might have produced, it seems we can go further and say that c1 is not just *permissible* but also *obligatory*. But that's just to say that c2 is itself wrong.

But that result isn't at all consistent with the picture that EVNPBI itself provides. EVNPBI implies that f1 *isn't* better than f3 and that f2 *isn't* better than f4. Those facts would seem to support the position that c2 is permissible.

Now, the fact that EVNPBI supports the result that c2 is wrong isn't, on its own, a problem for EVNPBI. We have agreed only that it's highly plausible that c1 is permissible. But we haven't yet said anything about the plausibility of the claim that c2 is wrong.

Nonetheless, a question remains: will a plausible, more complete person-based view—one that takes probabilities into account; one that recognizes that the better chance can, in some sense, make things morally better—be able to avoid the result that f3 is worse than f1 and that f4 is worse than f2? Will it be able to avoid the result c1 is wrong?

I don't see how it can on an *affect-the-future* view. But things change if we approach the problem from the perspective of the *affect-just-the-choice* view. On that view, what the probabilities affect is the evaluation of the *choice* and the choice *alone*; they leave untouched how the futures themselves are to be ranked.[21]

References

Broome, John 1992. *Counting the Cost of Global Warming*. White Horse Press.

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

Broome, John 2015. "General and Personal Good: Harsanyi's Contribution to the Theory of Value." In Iwao Hirose and Jonas Olson (eds.), *The Oxford Handbook of Value Theory*: 249–266.

Feldman, Fred 1986. *Doing the Best We Can*. Reidel.

Garson, James W. 2013. *Modal Logic for Philosophers* (2nd ed.). Cambridge University Press.

Hare, Caspar 2007. "Voices from Another World: Must We Respect the Interests of People Who Do Not, and Will Never, Exist?" *Ethics* 117: 498–523.

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

Narveson, Jan 1973. "Moral Problems of Population." In Michael D. Bayles (ed.), *Ethics and Population*. Cambridge, Mass.: Schenkman.

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

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

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

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

Roberts, Melinda A. 2011a. "The Asymmetry: A Solution," *Theoria* 77: 333–367.

[21] An earlier version of this paper was presented at Workshop on Sustainability and Future People (Umeå University, February 13-14, 2020). I am very grateful to the participants of that workshop for their comments on that paper.

Roberts, Melinda A. 2011b. “An Asymmetry in the Ethics of Procreation,” *Philosophy Compass* 6/11: 765–776.

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

Anders Herlitz[1]

Fixing Person-Based Stakes in Distributive Theory[2]

This paper explores an often overlooked distinction in distributive theory and its importance. The paper illustrates that there are at least three ways to interpret substantive proposals of how to distribute goods based on what is at stake for different individuals (or their “claims” or “complaints”) and that the interpretation affects what the proposals recommend. It then argues that each of the interpretations is associated with significant problems since they all seem to violate plausible requirements of rationality. A fourth interpretation of how to understand person-based stakes is introduced, but some questions regarding whether this is compatible with distributive theories that appeal to such stakes are raised.

[1] Institute for Futures Studies, anders.herlitz@iffs.se

[2] Financial support by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences) is gratefully acknowledged.

1. Introduction

It is often thought that when different distributions of goods are evaluated one ought to evaluate them in terms of how well they cater to different individuals' valid interests, what I will in this paper refer to person-based *stakes*. What is at stake for different individuals is often framed in terms of *complaints* or *claims*. In the Scanlonian contractualist tradition, it is common to evaluate distributions in terms of the size of the complaints that individuals can charge against them or against the principles that would justify the distribution (Scanlon 1982, 1998; Reibetanz Moreau 1998). In Fred Feldman's desertarianism, "desert claims" are the basis for distributive judgments (Feldman 2016: ch. 2). One version of prioritarianism frames this theory in terms of claims: "people have stronger claims to receive some benefit the worse off these people are" (Parfit 2012: 437; see also: Adler 2012: 321–337). When Harry Frankfurt introduced the idea that inequalities between people who are sufficiently well off might not matter at all (what Paula Casal has called the "negative thesis" of sufficientarianism, Casal 2007: 299-303) he framed it in terms of claims: "It is possible for those who are worse off not to have more urgent needs or claims than those who are better off because it is possible for them to have no urgent needs or claims at all" (Frankfurt 1987: 35). But how are these stakes supposed to be fixed across possible worlds?

This paper highlights and discusses the importance of an often-overlooked distinction pertaining to distributive theories that make recommendations based on what is at stake for different individuals. The paper first shows that one can interpret distributive theories that make recommendations based on what is at stake for different individuals in at least three seemingly plausible ways and illustrates how these interpretations have different evaluative implications. Roughly put, the paper outlines how one might think about what is at stake for an individual when a certain option, O, is evaluated (i) in terms of how well off she is in O compared to how well off she is if nothing is done, (ii) in terms of how well off she is in O compared to all available alternatives, or (iii) in terms of how well off she is in O compared to how well of she is in each of the available alternatives. It then presents and discusses some challenges that arise for the different interpretations. A fourth interpretation that avoids the problems with the other three is finally introduced, but so are some questions regarding whether this interpretation is compatible with distributive theories that appeal to person-based stakes are raised.

One of the reasons the different interpretations of how to determine what is at stake for different individuals is rarely discussed might be that distributive theorists tend to focus on, and discuss, cases in which it does not seem to matter which interpretation one has in mind. Consider, for instance, Thomas Nagel's famous case which Derek Parfit used to introduce prioritarianism (Parfit 1991: 1) and which Frankfurt discusses when he introduced sufficientarianism (Frankfurt 1987: 36):

Suppose I have two children, one of which is normal and quite happy, and the other of which suffers from a painful handicap. Call them respectively the first child and the second child. I am about to change jobs. Suppose I must decide between moving to an expensive city where the second child can receive special medical treatment and schooling, but where the family's standard of living will be lower and the neighborhood will be unpleasant and dangerous for the first child – or else moving to a pleasant semi-rural suburb where the first child, who has a special interest in sports and nature, can have a free and agreeable life (Nagel 1979: 123–124).

In this situation, it does not matter how one determines what is at stake for the different children. More is at stake for the second child regardless of whether one determines the person-based stakes by comparing moving to the city with each of the available alternatives, with available alternatives, or with how well off the second child is if nothing is done. Similar convergence of the different interpretations of what is at stake for different individuals will appear any time only two options are compared to each other. However, as I will describe in the following section, in richer choice situations, the different interpretations come apart.

2. Stakes

Many distributive theorists appeal to what is at stake for different individuals to describe and defend their theories. Examples are found both among those who think distributive theory should primarily ground betterness rankings of outcomes and those who think distributive theory should primarily tell us what distributors ought to do. Some go as far as suggesting that person-based stakes are the very object of fairness:

> Claims, and not other reasons, are the object of fairness. Fairness is concerned only with mediating between the claims of different people. If there are reasons why a person should have a commodity, but she does not get it, no unfairness is done her unless she has a claim to it (Broome 1991: 195).

Roughly, the general idea for teleologically-minded distributive theorists is that studying what is at stake for different individuals in different outcomes tells one how to rank the outcomes, whereas the general idea for those who promote and discuss deontic distributive theories is that one ought to act in a way that is supported by the fact that one takes what is at stake for different people into account in an appropriate way. How person-based stakes matter differs across distributive theories. Some aggregate all of them, some aggregate some of them, and some focus only on the strongest one.

Although stake-based distributive theories do not always provide precise accounts of why they emphasize the importance of person-based stakes, some general underlying motivations can be observed. What is at stake for different individuals is often ascribed importance because it is seen as a way of respecting individuals as persons (Scanlon 1982, 1998). Stake-based theories thus have a sense of person-centeredness which theories that merely focus on welfare levels in different outcomes seem to lack. Sometimes, this is associated with so-called second-personal ethics which suggests that ethical principles ought to be grounded in the second-personal relation that people who live together stand in with respect to each other (Darwall 2006). When I accidentally step on someone's foot, something is relevantly at stake *for her*. *She* – the person in front of me, who I can address in second-personal terms – has a valid complaint against *me* – the person who faces her – and can justifiably ask me to remove my foot. Based on how we as individuals stand in these second-personal relations to each other, we can develop distributive principles that reflect what we as individuals who relate to each other in second-personal ways can justifiably ask of each other (Darwall 2006: ch 12). Within approaches that aspire to respect individuals as persons and/or are based on second-personal relations, it is natural to focus on individuals' claims and complaints, what is at stake for different people. What is at stake for different individuals reflect what each individual brings to the table, so to speak.

Consider two typical presentations of stake-based approaches in distributive theory. First, Sophia Reibetanz Moreau's Minimax Complaints Model, which is an often referenced interpretation of Thomas Scanlon's contractualist view (see Scanlon 1998, 2013; Parfit 2003; Frick 2015; Kumar 2015; Suikkanen 2019; Bognar & Herlitz MS):

> An individual can reasonably reject a principle if her level of well-being and burden, given widespread acceptance of that principle over her lifetime, combine into a complaint greater than that had by anyone else about some alternative principle, given widespread acceptance of that alternative over a lifetime (Reibetanz Moreau 1998: 300).

Second, Alex Voorhoeve's Aggregate Relevant Claims (which can be seen as a general description of claim prioritarianism if one sets aside point 4):

1. Each individual whose well-being is at stake has a claim on you to be helped. (An individual for whom nothing is at stake does not have a claim.)
2. Individuals' claims *compete* just in case they cannot be jointly satisfied.
3. An individual's claim is *stronger*:
    a. the more her well-being would be increased by being aided; and
    b. the lower the level of well-being from which this increase would take place.
4. A claim is *relevant* if and only if it is sufficiently strong relative to the strongest competing claim.
5. You should choose an alternative that satisfies the greatest sum of strength-weighted, relevant claims (Voorhoeve 2014: 66).

These views, which are clearly different in their substantive recommendations, share a subtle but important inconclusiveness: they leave it unspecified how one ought to understand and measure what is at stake for different individuals (i.e. the complaints/claims) when one uses the theories to rank outcomes and/or to determine what one ought to do. More precisely, they say nothing about which possible worlds matter for understanding what is at stake for different individuals when one uses the theories to rank outcomes and/or determine what one ought to do. To see this, consider three interpretations of how to rank options with respect to what is at stake for different individuals.

> **The input view**: What is at stake for an individual when an option is evaluated is determined by comparing the individual's circumstances in that option and their circumstances such as they are if nothing is done.

On this view, what is at stake for the two children in Nagel's example when moving to the city is evaluated should be determined by comparing how well off the children would be if nothing is done and how well off they would be if the family moves to the city (and the person-based stakes associated with moving to the suburb should likewise be determined by comparing how well off the children would be if nothing is done and how well off they would be if the family moves to the suburb). This is the most literal reading of the quotes above. When Reibetanz Moreau refers to "an individual" and what complaints she might have toward different principles, it is natural to interpret this as a reference to an individual and their circumstances in the world such as it is unless something is done. When Voorhoeve describes how the strength of a claim depends on the "well-being from which an increase would take place," it is natural to interpret this as referring to the state of the world such as it is if there is no increase, no intervention. Generally speaking, the input view seems plausible if one wishes to ground distributive views in respect for persons such as they are.

> **The global view:** What is at stake for an individual when an option is evaluated is determined by comparing how well off the individual is in that option and how well off he or she is in all other options.[3]

On this view, what is at stake for the two children in Nagel's example when moving to the city is evaluated should be determined by comparing how well off the children would be in all other options (and likewise for the suburb). This is a less literal reading of the descriptions of stake-based distributive views. However, if one thinks of "an individual" in a more abstract way one can interpret Reibetanz Moreau's view in these terms. Furthermore, in light of how Voorhoeve determines what claims are relevant by appealing to what could be the case in some possible outcome, he is at least committed to making evaluations of what claims are relevant by considering how well off the best off person is in the outcome in which she is best off. This view reflects ascribing importance to what might be the case and expresses respect for the possibilities for each person.

> **The binary view**: What is at stake for an individual when two options are compared is determined by comparing how well off the individual is in the two options.4

On this view, one ranks options by making pairwise comparisons of all alternatives. What is at stake for the two children in Nagel's example when moving to the city is compared to moving to the suburb should be determined by comparing how well off the children would be in each of options. This interpretation of how to fix claims might seem peculiar from the perspective of philosophical intuition and standard vernacular, but it is plausible in light of social choice theory and the Condorcet tradition which focuses on pairwise comparisons. There are thus some technical reasons that might lead one to embrace this interpretation. Furthermore, and perhaps for this reason, it is an interpretation that Matthew Adler – one of few distributive theorists who makes it clear how he interprets person-based stakes (claims) – has defended (Adler 2012: 331).

[3] Cp. Fleurbaey et al. 2009.

[4] Cp. Fleurbaey et al. 2009

Various substantive normative views can be interpreted in accordance with each of these views. In the Minimax Complaints Model, are complaints against an option fixed by comparing how well off individuals are in the option to how well off they are if one does nothing, by comparing how well off individuals are in the option and how well off they are in the option in which they are best off, or for each pairwise comparison of options by looking only at the features of *these* options? In Aggregate Relevant Claims, are claims to benefits fixed by comparing how well off individuals would be if they received the benefits to how well off they are if nothing is done, by comparing how well off individuals would be if they received the benefits and how well off they are in the option that is the worst for them, or for each pairwise comparison of options by looking only at the features of *these* options?

For many substantive normative views, the interpretation of how one fixes and measures person-based stakes will have significant implications when the view is put into practice. Not in all cases, but in some cases. Consider an illustration. Assume that one is concerned only with the distribution of well-being and that well-being is the appropriate currency of person-based stakes. And assume further that one wishes to evaluate three principles (A, B and C) with respect to the Minimax Complaints Model. How should the following options be evaluated? (Each number describes the lifetime well-being of an individual if the principle is widely implemented.)[5]

A	B	C
[4, 9, 16]	[9, 16, 4]	[16, 4, 9]

On the input version of the Minimax Complaints Model, these options should be evaluated based on the complaints that can be charged against the options such that the strength of the complaints charged against an option is determined by comparing how well off individuals are in that option and how well of they are in the option that reflects doing nothing, the world such as it is in case no distributor intervenes in it. This means that in the case above, the ranking is entirely dependent on which of these options is the option that represents not intervening in the world (the *input*). If A is the input, there is no complaint against A, the strongest complaint against B is charged by the third individual who would go from well-being level 16 to 4, and the strongest complaint against C is charged by the third individual again, who in that option would go from 16 to 9. The options are thus ranked A > C > B (where ">" denotes *better than*). If B is the input, following the same reasoning, the options are ranked: B > A > C. And if C is the input, the ranking is C > B > A.

[5] I am indebted to Daniel Ramöller for helping me working out the details of this illustration.

On the global version of the Minimax Complaints Model, the ranking is based on the complaints that can be charged against the options such that the strength of the complaints charged against an option is determined by comparing how well off individuals are in that option and how well off they are in the option in which they are best off. In the case above, this gives the following ranking: A = B = C (where "=" denotes *equally as good as*), because in each option there is an individual at well-being level 4, and it is true for that individual that there is some option in which his or her well-being level is 16 (in A, the first individual has well-being level 4, and he or she could have had 16 (C); in B, the third individual has well-being level 4, and he or she could have had 16 (A); and in C the second individual has well-being level 4, and he or she could have had 16 (B)). This is the same ranking as the one provided by theories that only focus on the distribution of well-being in the different options and which makes no reference to claims at all (e.g., certain prioritarian views that rank outcomes "according to the sum of a strictly increasing and strictly concave transformation of well-being numbers", Adler & Holtug 2019: 3). However, it is not generally true that these views are coextensive. To see this, consider a contracted set of options where A is no longer a member. On the global version of the Minimax Complaints Model, B > C when only B and C are compared (the strongest complaint against B is charged by the first individual, who is at well-being level 9 in B and 16 in C, whereas the strongest complaint against C is charged by the second individual who is at well-being level 4 in C and 16 in B; the largest complaint against C is greater than the largest complaint against B).[6]

On the binary version of the Minimax Complaints Model, the ranking is established by pairwise comparisons of the options. In each pairwise comparison, there is some individual who in one of the options has well-being 4 and in the other has well-being level 16 (when A and B is compared, it is the third individual; when B and C are compared, it is the second individual; and when A and C are compared, it is the first individual). This is the largest complaint in each pairwise comparison of options, and the option in which this individual has well-being level 16 is the better one in each comparison. Since it is different individuals who have these complaints in each pairwise comparison, the following cyclical ranking of the options follows: B > A > C > B.

The purpose of the example above has been to illustrate the fact that how one fixes person-based stakes can have implications for how one evaluates different options. Although I have used the Minimax Complaints Model to illustrate the importance of how one interprets person-based stakes above, similar results arise for

[6] It is, however, generally true that the global view is coextensive with the binary view in all cases where there are only two options.

several other theories. Yet, it should be noted that the exact nature of how the recommendations of a theory depends on how one fixes person-based stakes cannot be entirely separated from the substantive details of the distributive theory. For instance, for claim prioritarianism (and for the parts of Voorhoeve's Aggregate Relevant Claims model which are prioritarian) it depends entirely on how one specifies and applies the "priority weights", the premium ascribed to claims of the worse off. There is a specific way of thinking about priority weights for which it is true that the views are coextensive.[7] However, there are other – perhaps more intuitive – ways to apply strictly increasing and strictly concave priority-weights-functions that generate different results depending on how one fixes claims.[8] Since there are these differences between distributive theories, it is important for proponents of stakes-based distributive theories to explore what the implications of applying different interpretations of how to fix person-based stakes are for their distributive theory.

3. Challenges

The previous section illustrated how it sometimes matters how one determines person-based stakes for evaluative purposes. Whatever reason one might have for

[7] The views are coextensive if the priority weights are applied to all well-being levels before the claims are aggregated. For instance, let the square root function (√) be the function that expresses the priority weights. When evaluating the outcomes in A, B and C, apply √ to the well-being levels (4, 9 and 16) before ranking the options, i.e. transform them to 2, 3, 4. On the binary view, A = B, B = C and A = C (because in each pairwise comparison one individual will have a priority-weighted claim equal to 2 to one option and two individuals will have a priority-weighted claim equal to 1 to the other option. On the global view, A = B = C (because for each option it is true that one individual has a claim equal to 2 to it, one individual has a claim equal to 1 to it, and one individual has no claim to it). On the input view, A = B = C (because regardless of what the input is, the claims to the input option will be 0, and the claims to each of the other views will also be 0, since for one option one individual will have a claim equal to 2 and the other individuals will each have a claim equal to -1 and for the other option one individual will have a claim equal to -2 and the other individuals will each have a claim equal to 1). Since the claims within options view satisfies the permutation axiom it is obvious that it ranks the options in the same way.

[8] Here is a simple illustration of how the binary view and the global view might differ from other prioritarian views, e.g. the view Adler and Holtug describe (Adler & Holtug 2019). Claims of individuals who are at well-being level 4 are weighted by 7. Claims of individuals who are at well-being level 9 are weighted by 4. Claims of individuals who are at well-being level 16 are weighted by 0.5. The function is strictly increasing and strictly concave. Compare the following two options: A and B from above, i.e. {4, 9, 16} and {9, 16, 4}. On Adler and Holtug's prioritarian view, the options are equally good. On the binary view and the global view: A > B (because the third individual's claim to A is 12 X 7 and the negative claims of the other individuals are 5 X 4 + 7 X 0.5, which gives a total priority-weighted claim to A = 60.5, while the priority-weighted claim to B = 5 X 7 + 7 X 4 – 12 X 0.5 = 57). But also the choice of whether to go for the binary view and the global view might matter for prioritarians. To see this, use the same priority weights as above and add that claims of individuals who are at well-being level 1 are weighted by 50. Apply these priority weights and evaluate the following three options: A = {4, 9, 16}, B = {9, 16, 4} and Z = {4, 1, 4}. Whereas the binary view provides the following ranking: A > B > Z, the global view gives the following ranking: B > A > Z (because on the global view the priority-weighted claims to A is merely 0 + 8 X 50 + 12 X 7 = 484, whereas the priority-weighted claims to B is 5 X 7 + 15 X 50 + 0 = 785.

espousing a stake-based approach to distributive theory, there are various formal challenges that arise for different interpretations. This section introduces some such challenges. To some extent, also the challenges depend on the substantive nature of the distributive theory. However, they merit a general discussion because stake-based approaches to distributive theory are so common, and it is important to know what problems they face.

3.1 The input view, Normative Invariance, Satisfiability and the Stability Condition

The input view faces a wide range of challenges. For instance, ascribing substantial importance to the input world – the world such as it is if one does not intervene in it – requires that one knows what the input world actually is, and sometimes that is not obvious: is continuing an ongoing treatment plan not to intervene or is it an intervention? Furthermore, views that base their recommendations on how well off people are in the world such as it is if one does not intervene in it have difficulties informing choice when the population changes, i.e. when some people will come into existence if one acts in one way and other people will come into existence if one acts in a different way (Herlitz & Eyal MS).

However, input interpretations of stake-based views also face more formal challenges. First, they appear to violate Normative Invariance (see Carlson 1995; Bykvist 2007; Arrhenius & Rabinowicz 2014):

> **Normative Invariance:** An action's normative status does not depend on whether or not it is performed (Bykvist 2007: 99).

Consider again the distribution from the previous section and what claim prioritarians who adopt the input view would say:

A	B	C
[4, 9, 16]	[9, 16, 4]	[16, 4, 9]

If A is the input, a prioritarian input view implies that one ought to choose C, yet once one has made that decision – performed the choice act – the input changes to C, and when C is the input choosing C over A is impermissible (since input prioritarianism ranks the options in the following way when C is the input: B > A > C).

However, since input views ascribe significant importance to what the world looks like if nothing is done, it might be plausible to accept that making a choice has implications for how alternatives ought to be individuated (see Herlitz 2020a). Perhaps choosing C when A is the input is not the same as sticking with C when C is the input. If that were the case, Normative Invariance would not be formally violated.

Second, they seem to violate Satisfiability:

> **Satisfiability**: For any agent and any possible situation, there is an action such that if the agent were to perform the action in this situation, then she would conform to the theory (Bykvist 2007: 116).

If a distributive theory tells one to choose C, but implies that when one has chosen it A is better, has one "conformed to the theory"? This seems doubtful.

Third, consider the following condition:

> **The Stability Condition**: A decision method/normative theory, P, meets the stability condition if and only if it is always true according to the method/theory that if an option, X, that according to the method/theory is maximal (i.e. not worse than any alternative) in a set of alternatives, C, is chosen, then the transmutation$_X$ of X, X_X, is also maximal according P in C_X, the set of alternatives consisting of the transmuted$_X$ alternatives in C (Herlitz 2020a: 6).

Transmutation$_X$ is defined as follows:

> **Transmutation$_X$**: A transmutation$_X$ of an alternative, Y, in a set of alternatives, C, of which both X and Y are elements, into a transmuted$_X$ alternative, Y_X, is the transmutation of Y that appears in the choice set, C_X, that is the set of alternatives C in which the negative and positive values associated with choosing X have been dispersed across the alternatives in C (Herlitz 2020a: 5).

In the example above, following the input-prioritarian principle and choosing C from {A, B, C} turns C into the transmuted$_C$ C: C_C. Since C in {A, B, C} is a maximal alternative according to the input-prioritarian principle and C_C in {A_C, B_C, C_C} is not a maximal alternative according to the input-prioritarian principle this principle violates the Stability Condition.

Following a principle that violates the Stability Condition implies that one's choices are not stable; they fail to remain robustly supported by the principle after they have been made. Consider an illustration of what it means to violate this condition. Imagine someone walking into a pizza parlor bombastically declaring his order: "I'll have the Margarita, unless I've already ordered that, in which case I'll go for the Napoletana! Although, were you to think that was my order I'll order the Margarita!" This behavior is strikingly irrational.

3.2. The global view, Basic Contraction Consistency and "irrelevant" alternatives

As the attentive reader might have already noted, the global view seems to violate what many take to be a basic requirement of rationality, so-called *Basic Contraction Consistency* (sometimes called "alpha" or "the Chernoff Condition"; see Chernoff 1954; Sen 1970, 1993; Fleurbaey et al. 2009; Voorhoeve 2014; Herlitz 2019a, 2020a):

> **Basic Contraction Consistency**: If an option, X, is permissible in a set of alternatives, it is also permissible in a subset of these options containing X.

As pointed out in the previous section, if the Minimax Complaints Model is interpreted in line with the global view it states that A, B and C are all permissible if the set of alternatives is {A, B, C}, while only C is permissible if the set of alternatives is {B, C}. {B, C} is a subset of {A, B, C} and B is a member of both sets, but although B is permissible in the larger set it is not permissible in the contracted, smaller set. Basic Contraction Consistency thereby seems to be violated.

A somewhat common response to the accusation that one's view violates Basic Contraction Consistency has been that the argument fails to take into account the importance of individuating the alternatives appropriately (see Sen 1993; Arrhenius 2009; Voorhoeve 2014; Herlitz 2019a, 2020a; Brown 2019). If the alternatives change together with the changes in the choice set, they ought to be differently individuated, and with a different individuation there is no violation of Basic Contraction Consistency. The fact that A, B and C are all permissible in the choice set {A, B, C} but only C* is permissible in the choice set {B*, C*} does not show that Basic Contraction Consistency has been violated. It is true that B is permissible in the first set and B* is not permissible in the second set, but B and B* are not the same option, so the second set does not contain an option that was permissible in the initial set (and {B*, C*} is not a subset of {A, B, C}).

The extent to which this response is successful partially depends on what status one ascribes to rationality axioms such as Basic Contraction Consistency and partially on what substantive distributive theory one evaluates. As John Broome has pointed out, of course all options can in principle be individuated differently in all instances, but if one takes such a radical approach, it is meaningless to even discuss requirements of rationality (Broome 1991: ch. 5). One must thus ask: when *ought* one change the individuation of an alternative?

A common answer to that question is that one ought to individuate alternatives so that the features that are pertinent grounds for choice in light of the principle one uses are included in the individuation (Broome 1991: ch. 5; Voorhoeve 2014; Herlitz 2020a, 2019a). On the global view the option in which an individual is worst off partially determines what one ought to choose. If that changes when the set of options is contracted, it seems reasonable to accept that also the individuation ought to change (see Voorhoeve 2014). This means that Basic Contraction Consistency is not violated.[9]

But there is a different, somewhat similar, challenge for the global view: it ascribes a peculiar amount of importance to options that at an intuitive level are obviously irrelevant. Consider the case from the previous section again:

A	B	C
[4, 9, 16]	[9, 16, 4]	[16, 4, 9]

If {A, B, C} is the entire choice set (i.e. all options available), the global view deems A, B and C to be equally good options. However, if there were *more* options in the choice set, this is not necessarily true any longer. To see this, consider the same case with an added option, D, which is by no means an appealing option, and no serious normative view would suggest anyone to choose it, but which still has significant importance according to the global view:

A	B	C	D
[4, 9, 16]	[9, 16, 4]	[16, 4, 9]	[1, 2, 2]

Everyone is worse off if D is chosen, and no serious stake-based distributive theory suggests that one ought to choose D (regardless of how one interprets how person-based stakes are fixed). Nevertheless, the fact that the first individual is worse off than the others in D has implications for how the global view evaluates the *other*

[9] The fact that Basic Contraction Consistency is not violated because one is warranted in changing the individuation of the options solves the formal problem this argument points toward, but there might remain practical problems. For instance, changing the individuation of the alternatives does not undermine the possibility of making money-pump arguments, so if such arguments have any bearing on what distributive theory is plausible the global view can be challenged on such grounds (see Herlitz 2019a).

options. Consider, for instance, stake-based theories that ascribe greater weight to benefits to the worse off. If person-based stakes are fixed with the global view, those theories imply that C is the best option, because compared to D (the option in which everyone is worst off in this example), C is the option with most benefits to the worst off (the first individual).[10]

To grasp how counterintuitive this implication is, consider an illustration. Imagine that I have two equally good students, Aya and Monika, who are both on the job market competing for the same jobs. Both of them ask me to help them. I am impressed by both of them and happy to provide them with guidance and whatever reference letters can help. I also do not play favorites when I write these letters, because I find it really hard to rank them and I truly think they are equally good. One day, I learn that my close friend Emmanuelle chairs the hiring committee at a department that would suit both Aya and Monika perfectly. I know Emmanuelle well, so I am confident that if I call her and push for one of my students, this will significantly increase her chances, but if I call and try to push for both of them, Emmanuelle will not take me seriously, so it will have no impact at all on their prospects. I must choose: should I call Emmanuelle and push for Aya or call and push for Monika? I could of course also decide to not call Emmanuelle at all, or to call and push for both my students. The latter options have identical implications for Aya and Monika. I determine that calling to push for Aya and calling to push for Monika are the two best options, and they are equally good. Aya and Monika have identical person-based stakes in the situation. The best I can do is to randomize, flip a coin. But then it strikes me: many, many years ago, Aya expressed fascination for and had an interest in writing about Friedrich Hayek. I know Emmanuelle *hates* Hayek. So, if I were to call Emmanuelle in order to try to make my students' chances *worse,* I would be more successful at discouraging Emmanuelle from hiring Aya than I would be at discouraging her from hiring Monika. If I compare how badly off Aya is in the option that is worst for her and how badly off Monika is in the option that is worst for her, it is obvious that Aya is worse off. That settles it! Since I can ruin Aya's life more effectively than Monika's, I ought to call Emmanuelle and promote Aya. This is the radically implausible implication of the global view combined with priority to the worse off.

There are some recent attempts at dealing with cases like this by qualifying

[10] It could be objected that egalitarianism is a serious distributive theory and that egalitarianism implies that D is the best of all options. It is true that an extreme egalitarian would favor D over the other options, but most egalitarians reject the idea that one ought to choose an option in which everyone is made worse off and say that although equality matters it is not all that matters (e.g. Temkin 2003). Furthermore, one can construe examples where not even an extreme egalitarian would be attracted by the option that the global view ascribes importance in this way. For instance, should one choose E, F, G or H, where E = {10, 9, 9}, F = {9, 10 9}, G = {9, 9, 10} and H = {1, 5, 5}? Is it plausible to think that E is the best option in virtue of the fact that H is available in this case?

which options are allowed to matter when one evaluates a set of options. For instance, Michael Otsuka has in a recent paper suggested that certain options should be excluded from the set of options that influence one's evaluations (Otsuka 2017). As he puts it: when making distributive decisions, one ought to only care about "genuine moral options", options that are not "manifestly unreasonable" (Otsuka 2017: 200). Intuitively, neither D nor calling Emmanuelle in order to ruin Aya's chances to get hired are what Otsuka calls "genuine moral options", and most would agree that they are in fact "manifestly unreasonable".

However, Otsuka's proposal is quite unsatisfactory in that it is very imprecise and relies merely on intuition. The problem is structural and solutions that rely on vague notions such as "genuine moral options" and "manifestly unreasonable" are unsatisfactory. On the one hand, there will be a plethora of cases in which it is not clear whether an option should be considered a genuine moral option or not. On the other hand, one wants an explanation of what makes something a genuine moral option, which Otsuka does not provide.

Perhaps there are ways of making views like Otsuka's more precise, but this is no easy task. For instance, one could define "genuine moral options" in terms of options that are not Pareto dominated, i.e. options for which it is true that there is no alternative in which someone is better off but no one is worse off. That would certainly rule out D from the set of options that are allowed to influence one's evaluation in the example above. However, and here is why this kind of approach must be explored in much more depth than I can afford it in this paper: can we be sure that it would rule out calling Emmanuelle to ruin Aya's chances? No. If I decide to call Emmanuelle to ruin Aya's chances, I *de facto* improve the chances of all *other* applicants. Their chances will of course not improve by much but removing one competitor must still count as a benefit to them. This means that this option is not Pareto dominated. It might seem Pareto dominated by the option "do not call Emmanuelle at all", but the option is in fact worse for the other candidates, so it is not Pareto dominated.

3.3 The binary view and cycles

As noted already above, the binary view occasionally generates cyclical rankings of options (see Norcross 2002; Parfit 2003; Fleurbaey et al. 2009). The Minimax Complaints Model together with the understanding that complaints should be fixed according to the binary view ranked the three options as follows: A > B > C > A. This means that the better than relation (i.e. >) that ranks the options violates transitivity:

> **Transitivity**: *R* is transitive if and only if, for all *a*, *b* and *c*, if *aRb* and *bRc*, then *aRc*.

All conventional positive comparative relations (i.e. *more F than*, *less F than*, *equally as F as*) are usually considered to be transitive.[11] And accepting a principle that implies that a conventional positive comparative relation violates transitivity is by many considered irrational.

Although there are exceptions (most notable in the contemporary philosophical literature: Larry Temkin (1996, 2012)), few are willing to accept that a normative principle that generates cyclical rankings (and implies intransitive better than-relations) is even remotely plausible. As Broome puts it, giving expression to a very strong view on the matter:

> A comparative relation is necessarily transitive. This is an analytic feature of the operator 'more…than': the meaning of 'more…than' implies that 'more F than' is transitive. The more formal features of the meaning of a term are often called the 'logic' of the term. Deontic logic is the logic of 'ought', for example. In this sense, transitivity is a feature of the logic of 'more…than' (Broome 2004: 50).

Since the binary view together with certain substantive views of how to evaluate person-based stakes generate cyclical evaluations, this view seems to come with a significant shortcoming: one must abandon what Broome calls "the logic of 'more… than'".

Nevertheless, it is worth pointing out that there might be ways of making this position less unpalatable. One problem with accepting cyclical evaluations is that principles that generate cyclical rankings of options seem unable to guide choice: whatever one chooses, there is an option that was ranked above it (see Handfield 2015). According to standard rational choice theory that says that it is irrational to choose an option that is worse than some alternative, rational choice seems impossible. More strongly, some believe that cyclical rankings not only undermine the possibility of rational choice, but the possibility of practical reasoning as such:

> If transitivity fails, then it may seem impossible for the better than relation to guide our choices or attitudes in many cases. This is because, for any outcome we choose or prefer, there may be some better alternative that we ought to have chosen or preferred instead. The rejection of transitivity might, therefore, lead to

[11] This is, by contrast, not true for non-conventional positive comparative relations (e.g. "parity", "incommensurability", "imprecise equality") such as these have been defined and discussed in the recent value-theoretical literature (see Chang 2002; Rabinowicz 2008; Carlson 2010; Parfit 2016; Herlitz 2019a).

> skepticism about practical reasoning, or at least about the role of the good in practical reasoning (Nebel 2018: 875).

There is no space to present the details of these approaches, but it should at least be noted that there are alternative approaches to rational choice that can be used together with principles that generate cyclical rankings so that these principles preserve at least some element of being able to guide choice. For instance, approaches that suggest that a rational choice is a choice of an element in the so-called Schwartz Set and approaches that suggest that a rational choice is a choice of an element in the set of Strongly Uncovered options can both be combined with principles that generate cyclical rankings so that these principles preserve some element of action-guidingness.[12] Just as an illustration, imagine that a principle, X, ranks the options *a, b, c* and *d* in the following way:

a > b; b > c; c > a; a > d; b > d; c > d

X generates a cyclical ranking (of *a, b* and *c*) and there is no option that is not worse than any alternative with respect to X. Whether X is action-guiding can thus be questioned. However, only *a, b* and *c* are in the Schwartz Set, and only *a, b* and *c* are Strongly Uncovered, so if one uses any of these criteria for what it means to follow a principle one must admit that X has *some* action-guidingness; it says: do not choose *d*.

4. Person-based stakes within options?

It might seem, thus, as if proponents of stake-based approaches to distributive theory ought to reject the input view, the global view, and the binary view. There is, however, an alternative – and arguably far less intuitive – way to think of person-based stake:

> **The within options view**: What is at stake for an individual when an option is evaluated is determined *within each option,* and without any reference to any other possible option.

[12] The Schwartz Set is a subset of the set of options that consists of the alternatives in all Schwartz Subsets, i.e. the subsets of the set of options such that (1) every element of the subset is pairwise unbeaten by every non-element in the subset, and (2) there is no non-empty subset of the subset that satisfies the first condition (see Schwartz 1972, 1986; Herlitz 2020b, 2020c). An option is Strongly Uncovered if no alternative strongly covers it. An option, *x*, strongly covers an alternative, *y*, if and only if it is true that $x > y$ and compared to any other alternative, *z*, if $z > x$ then $z > y$, and if $y > z$ then $x > z$ (see Duggan 2013; Herlitz 2020b, 2020c).

Recall the case from above:

A	B	C
[4, 9, 16]	[9, 16, 4]	[16, 4, 9]

On the within options view, these options are equally good. The within options view fixes the size of each person-based stake only on the basis of the features each individual has in each option. The view thus satisfies what Broome calls “impartiality” and what Matthew Adler calls “anonymity” or the “permutation axiom”: it does not matter which individual has what features, only the pattern of the features matter (Broome 2004: 135; Adler 2012: 52). Since the pattern of the distribution of well-being is identical in the three options (one individual at level 4, one at level 9, and one at level 16), the options have the same normative status. This specification gives the following ranking: A = C = B.

On this view, there are no cyclical rankings, no violation of Basic Contraction Consistency, no importance is ascribed to seemingly irrelevant options, and there is no violation of the Stability Condition. However, one might wonder what a person-based stake that is established within an option and only there actually is. Person-based stakes such as claims and complaints refer to the absence of something; they are directed toward something that is not there. What could that be if one accepts the within options view?

Whereas there are views that are naturally understood as within options views,[13] it seems one must resort to quite unconventional and rather creative interpretations of “claims” and “complaints” to combine other distributive theories with the within options views. Claims and complaints are naturally understood in *comparative* terms. They make reference to different states or positions that an individual can inhabit. In the most obvious case: the state that an individual is in and the state that she ought to be in. For egalitarians, the relevant comparisons are the comparisons between how well off different individuals are, and it is obvious that one very plausible interpretation of egalitarianism is that what matters is how well off different individuals are within one and the same option. Other distributive theories do not ascribe importance to comparisons between individuals. In fact, other distributive theories often make it a virtue that they are not comparative in this sense. But what is then the comparative element of the person-based stakes? On the within options view, there is none.

[13] Temkin’s egalitarianism is naturally understood in such a way for example, see Temkin 1993; Adler 2012: 321-328. And sufficientarian theories that refer to sufficiency levels can easily be understood in these terms, see Frankfurt 1987; Casal 2007; Herlitz 2019e.

Consider again Broome's remarks regarding the transitivity of conventional comparative relations:

> A comparative relation is necessarily transitive. This is an analytic feature of the operator 'more...than': the meaning of 'more...than' implies that 'more F than' is transitive. The more formal features of the meaning of a term are often called the 'logic' of the term. Deontic logic is the logic of 'ought', for example. In this sense, transitivity is a feature of the logic of 'more...than' (Broome 2004: 50).

Broome suggests that transitivity is a feature of the logic of 'more...than', and those who suggest that better than is not a transitive relation fail to respect the logic of the term. A similar argument can be made against (certain) stakes-based approaches that suggest that person-based stakes are fixed within each option and without any reference to any other options.

Here is how David Gauthier presents the role that claims play in a bargaining situation:

> (4) A *claim* is a demand by a prospective co-operator for a particular co-operative surplus, made initially in the bargaining.
>
> (5) The *claim point* is the point in utility space representing the (possibly hypothetical) outcome that would afford each person a utility equal to that of his claim; it is feasible if and only if it is a point of the outcome-space (Gauthier 1986: 142).

Using Gauthier's language: what is the *claim point* if claims are fixed within an option and with no reference to any other option? For the Minimax Complaints Model, Aggregate Relevant Claims and claim prioritarianism, there cannot be any claim point at all if they are combined with the within options view.

But what is a claim if there is no claim point? What is a person-based stake if there is no claim/complaint to or against something? Analogous to Broome's point, it can be suggested that those who rely on conceptions of person-based stakes that have no claim points fail to respect the logic of these terms.

5. Discussion

This paper emphasized an ambiguity in stake-based approaches to distributive theory. A distributive theory that holds that distributive judgments ought to be based on what is at stake for different individuals in different options generates different recommendations depending on how these person-based stakes are fixed. I outlined three interpretations of how to understand how person-based stakes should be fixed and illustrated how the choice of interpretation has significant implications for what the contractualist Minimax Complaints Model, Voorhoeve's Aggregate Relevant Claims model and claim prioritarianism say. Although it is impossible to in a single paper present a general overview of how this issue matter for all stake-based approaches to distributive theory, I believe the arguments of this paper show that there is good reason to suspect that similar problems will arise for most if not all stake-based approaches. In other words, I hope to have shown the importance on being clear on what one means by person-based stakes (or claims/complaints) when these notions are used to describe a distributive theory.

In the second part of the paper, I presented a series of challenges that different interpretations of how person-based stakes should be fixed for evaluative purposes face. I showed that the view risks leading to cyclical evaluations, that the global view seems to violate Basic Contraction Consistency and also ascribe an unreasonable amount of importance to seemingly irrelevant alternatives, and that the input view fails to generate stable recommendations. I also introduced a view that seems to avoid the formal challenges: the within options view, but noted that this view faces what might be seen as an even greater problem: it fails to respect the logic of the person-based stakes.

This somewhat somber result gives one reason to paus and to question the usefulness of at all using person-based stakes and concepts such as claims and complaints in distributive theory. Why at all embark on the perilous path of presenting a stake-based view that run significant risk of violating basic requirements of rationality when one can formulate much simpler distributive theories that provide action-guiding recommendations based only on the internal features of the alternatives and which do not risk violating any rationality requirements?

The answer might lie in the desire to ground distributive theory in concerns for persons and their interests. If persons and their interests are what matters when we make distributive judgments, then we need a concept that describes different individuals' interests. Person-based stakes, claims and complaints are such concepts. But if these concepts are indeed unavoidable, we should also recognize the general challenge this actualizes: stake-based distributive theories violate basic requirements of rationality *unless they have certain substantive features*, i.e. unless they aggregate and weigh person-based stakes in specific ways. Proponents of stake-based distributive theories might in other words find themselves forced to find compromises between formal rationality requirements and well-grounded substantive distributive criteria. This is not the place to determine whether it is possible to develop and defend such compromises that remain attractive.

References

Adler, Matthew. 2012. *Well-being and Its Fair Distribution*. Oxford: Oxford University Press.

Adler, Matthew & Nils Holtug. 2019. "Prioritarianism: A response to critics." *Politics, Philosophy & Economics* DOI:10.1177/1470594X19828022

Arrhenius, Gustaf. 2009. "Can the person affecting restriction solve the problems in population ethics?" In *Harming Future Persons: Ethics, Genetics and the Nonidentity Problem*, ed. M. A. Roberts & D. T. Wasserman. Dordrecht: Springer.

Arrhenius, Gustaf & Wlodek Rabinowicz. 2014. "The value of existence." In *The Oxford Handbook of Value Theory*, ed. J. Olson & I. Hirose. Oxford: Oxford University Press.

Bognar, Greg & Anders Herlitz. MS. "Testing the decomposition test."

Broome, John. 1991. *Weighing Goods*. Oxford: Blackwell.

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

Brown, Campbell. 2019. "Is close enough good enough?" *Economics and Philosophy* DOI:10.1017/S0266267119000099

Bykvist, Krister. "Violations of normative invariance: some thoughts on shifty oughts." *Theoria* 77: 98–120.

Carlson, Erik. 1995. Consequentialism Reconsidered, Theory and Decision Library. Dordrecht: Kluwer.

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

Casal, Paula. 2007. "Why Sufficiency Is Not Enough." *Ethics* 117: 296–326.

Chang, Ruth. 2002. "The possibility of parity." *Ethics* 112: 659–688.

Chernoff, Herman. 1954. "Rational selection of decision functions." *Econometrica* 22: 422–443.

Darwall, Stephen. 2006. The Second-Personal Standpoint: Morality, Respect, and Accountability. Cambridge, MA: Harvard University Press.

Duggan, John. 2013. "Uncovered sets." *Social Choice and Welfare* 41: 489–535.

Fleurbaey, Marc, Tungodden, Bertil & Peter Vallentyne. 2009. "On the possibility of nonaggregative priority for the worst off." *Social Philosophy and Policy* 26: 258–285.

Frankfurt, Harry. 1987. “Equality as a Moral Ideal.” *Ethics* 98: 21–43.

Frick, Johann. 2015. “Contractualism and Social Risk.” *Philosophy and Public Affairs* 43: 175–223.

Gauthier, David. 1986. *Moral by Agreement*. Oxford: Clarendon Press.

Handfield, Toby. 2015. “Rational choice and the transitivity of betterness.” *Philosophy and Phenomenological Research* 89: 584–604.

Herlitz, Anders. 2019a. “Nondeterminacy, two-step models, and justified choice.” *Ethics* 129: 284–308.

Herlitz, Anders. 2019b. “The indispensability of sufficientarianism.” Critical Review of International Social and Political Philosophy 22: 929–942.

Herlitz, Anders 2020a. “Stable and unstable choices.” *Economics and Philosophy* 36: 113–125.

Herlitz, Anders. 2020b. “Non-transitive Better than Relations and Rational Choice.” *Philosophia* 48:179–189.

Herlitz, Anders. 2020c. “Correction to: Non-transitive Better than Relations and Rational Choice.” *Philosophia* 48: 431.

Herlitz, Anders & Nir Eyal. MS. “Input and output in distributive theory.”

Kumar, Rahul. 2015. “Risking and Wronging.” *Philosophy and Public Affairs* 43: 27–51.

Nagel, Thomas. 1979. “Equality”. In his *Mortal Questions*. Cambridge: Cambridge University Press.

Nebel, Jacob M. 2018. “The good, the bad, and the transitivity of *Better Than*.” *Noûs* 52: 874–899.

Norcross, Alastair. 2002. “Contractualism and Aggregation.” *Social Theory and Practice* 28: 303–314.

Otsuka, Michael. 2017. “How it makes a difference that one is worse off than one could have been.” *Politics, Philosophy and Economics* 17: 192–215.

Parfit, Derek. 1991. “Equality or Priority?” *The Lindley Lecture*. The University of Kansas.

Parfit, Derek. 2003. “Justifiability to each person.” *Ratio* 16: 368–390.

Parfit, Derek. 2012. “Another defence of the priority view.” Utilitas 24: 399–440.

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

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

Reibetanz Moreau, Sophia. 1998. "Contractualism and aggregation." *Ethics* 108: 296–311.

Scanlon, Thomas M. 1982. "Contractualism and utilitarianism." In *Utilitarianism and Beyond*, ed. A. Sen & B. Williams. Cambridge: Cambridge University Press.

Scanlon, Thomas M. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Schwartz, Thomas. 1972. "Rationality and the myth of the maximum." *Noûs* 6: 97–117.

Schwartz, Thomas. 1986. *The logic of collective choice*. New York: Columbia University Press.

Sen, Amartya. 1970. *Collective Choice and Social Welfare*. Amsterdam: North-Holland.

Sen, Amartya. 1993. "Internal consistency of choice." *Econometrica* 65: 745–779.

Suikkanen, Jussi. 2019. "Ex Ante and Ex Post Contractualism: A Synthesis." *The Journal of Ethics* 23: 77–98.

Temkin, Larry. 1996. "A continuum argument for intransitivity." *Philosophy and Public Affairs* 25: 175–210.

Temkin, Larry. 1993. *Inequality*. Oxford: Oxford University Press.

Temkin, Larry. 2003. "Egalitarianism defended." *Ethics* 113: 764–782.

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

Voorhoeve, Alex. 2014. "How Should We Aggregate Competing Claims?" *Ethics* 125: 64–87.

Studies on climate ethics and future generations, vol. 1
Working paper series 2019:1–11

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

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*