A Fairness-Based Astronomical Waste Argument

Patrick Krystof Kaczmarek

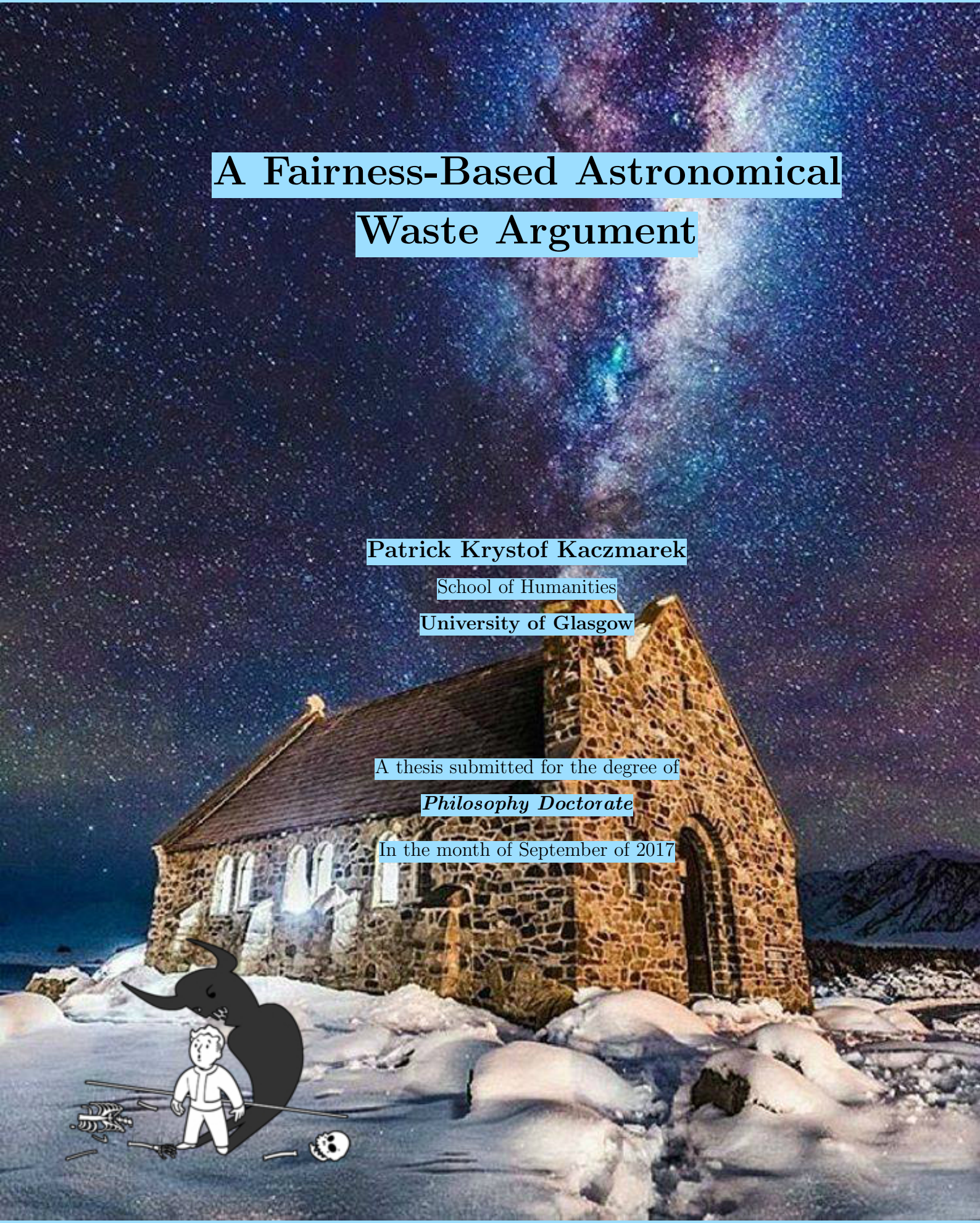
School of Humanities

University of Glasgow

A thesis submitted for the degree of

Philosophy Doctorate

In the month of September of 2017



1. **External Reviewer:**

Tim Mulgan

2. **Internal Reviewer:**

Stephan Kraemer

Day of the defense: November 22nd 2017

Signature from Chair:

Michael Brady

Abstract

I defend a modified version of Marc Fleurbaey and Alex Voorhoeve's *Competing Claims View* that captures an additional consideration of fairness in the context of variable populations. I call this consideration 'worthwhileness'. Part 1 goes on to argue that this view describes the expected value of a lottery in a way that is consistent with the axiological framework of Averagism. Also, I propose a novel definition of 'overpopulation', and explain why considerations of fairness so-described by Averagism support our other moral reasons for avoiding overpopulating the world. In part 2, I design and run a toy model to determine which development policy-option is best in terms of satisfying the Competing Claims View. One of these options is ambiguous insofar as it combines two intuitions which have time and again proven themselves rather difficult to jointly pin down. Putting them together forms what I will hereafter call, after its leading proponent, *Broome's Intuition About Neutrality* ('$\mathbb{BN}$'). I argue that there is at least one combination of a (mathematically) well-behaved axiology and bridge principle that yields a moral theory which satisfies the normative reading of $\mathbb{BN}$. Armed with all the right ingredients, we can now run the model. Based on some conservative assumptions, we find that we ought to take steps towards: (a) militating against the threat of a broken world; and (b) prolonging humankind's place in the stars (to some extent).

"What now matters most is how we respond to various risks to the survival of humanity. We are creating some of these risks, and discovering how we could respond to these and other risks. If we reduce these risks, and humanity survives the next few centuries, our descendants or successors could end these risks by spreading through this galaxy.

Life can be wonderful as well as terrible, and we shall increasingly have the power to make life good. Since human history may be only just beginning, we can expect that future humans, or supra-humans, may achieve some great goods that we cannot now even imagine. In Nietzsche's words, there has never been such a new dawn and clear horizon, and such an open sea.

If we are the only rational beings in the Universe, as some recent evidence suggests, it matters even more whether we shall have descendants or successors during the billions of years in which that would be possible. Some of our successors might live lives and create worlds that, though failing to justify past suffering, would give us all, including some of those who have suffered, reasons to be glad that the Universe exists".

Derek Parfit

This is for the late Derek Parfit. He was a towering giant in the field of population ethics, and his influence on my work (and life) is immeasurable.

Acknowledgements

Nick Bostrom and Daniel Dewey urged me to write on this topic while I was visiting the Future of Humanity Institute. I am indebted to them, as well as Eric Drexler, for the kind words and guidance. I'm also very glad that they do what they do so well. I owe a special debt of gratitude to my supervisors, Campbell Brown and Ben Colburn, for sticking it out with me. There were, to be sure, many occassions when I struggled or strayed from the right path. Ben especially went out of his way and has been, simply put, my guardian angel. I also learned a lot from Japa Pallikkathayil and Adam Rieger when they stepped in to supervise at different stages of the project. Michael Plant is owed a special thanks. He has been relentless in his criticism of my work (and life). He's my best friend, and I blame him entirely for any mistakes left in this dissertation. Very many more people contributed in some way or another to the success of this project. Too many for me to remember now, and I'm sure I would have left a few names out if I tried to list them all. You know who you are. I'm eternally grateful. Still, during the final stages of the dissertation a few persons really stepped up or got involved in some crucial way. For starters, what you hold in your hands is significantly better than what I had walking into my viva. Indeed, I am very glad to have had both Stephan Kraemer and Tim Mulgan perform as examiners. Their suggestions for improving the structure of the dissertation were especially helpful. *Again, I blame Michael for any shortcomings with respect to making Stephan and Tim's suggestions work.* Moreover, for helping me tease out my ideas or providing outstanding comments on /criticism of (sometimes rather large) portions of the dissertation, I wish to thank (in no particular order): Teru Thomas, Daniel Cohen, Simon Beard, Mikio Agaki, John Broome, Ben Levinstein, Hugh Lazenby, Julia Mosquera, and Derek Parfit.

I don't have the words to describe how glad I am for Jenifer 'Malka' Siegel playing her part in my life. Frankly, I'm not sure I would have submitted on time if at all had it not been for her. I am madly in love with you, Malka.

Oh yes, I mustn't forget:

"All hail Wiblin Hall!"

Contents

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

Listings

1

Précis

> For some minutes he lay there miserably, but when the five hundred and eight-seventh Heffalump was licking its jaws, and saying to itself "Very good honey this, I don't know when I've tasted better," Pooh could bear it no longer.
>
> A. A. Milne, *The House at Pooh Corner*

Our world is plagued with things that are very bad. Some of these bad things are happening right now. For example, disease and famine have both regularly featured in human history. Others can be seen slowly unraveling, such as the harms associated with overpopulation. Some of these bad things haven't yet happened and may never happen. All three categories involve threats to humanity's long-term potential.

We are creating some of these risks to humanity's survival. For example, there is some risk of a deadly pathogen being released accidentally, unilaterally by a misguided or misinformed agent acting from impersonal concern, or maliciously by some omnicidal group.[1,2] Furthermore, we are growingly becoming capable of, not only mixing deadlier virus-cocktails, but doing so outside of the laboratory. In both cases, governments and other relevant actors have been slow to react to this evolving landscape. This is distressing news; for if we now make poor choices, humanity may either go extinct or be irreversibly crippled. Things might get so bad that our progeny's lives are not worth living. Indeed, we might be responsible for having unleashed Hell on Earth.

There is no stage of humankind's history, going as far back as our cave-dwelling forebears, where we were safe. But this does not mean that the source of our problems has not changed. Unlike our forebears, who were incapable of much more than waging war on each other with heavy, blunt objects, and could have been snuffed out by a volcano (being less spread out, and few in numbers), we are far more likely to withstand a volcano (e.g.), but destroy mankind's long-term potential by our very own hands.

[1] (Bostrom et al., 2016b)

[2] For a typology of agential risks/terror see (Torres, 2016). The classic example of such bandits is the Aum Shinrikyo death-cult that released sarin gas in the Tokyo subway system in 1995.

1. PRÉCIS

A lot now hangs on the choices we make. Perhaps at some stage of history we might reach a level of maturity that insulates us from even man-made threats. But it won't be enough to merely survive the next few centuries.[3] There are infinitely many possible worlds in which we do survive in the long-run, spreading out into several galaxies, and yet life is worse, even far worse, than it would have been had we made some other choice along the way.[4]

Having said this, we must acknowledge the sour truth that our fate is at least partly out of our hands. Life could be terrible as well as wonderful *even if we took all the right steps*—e.g., a freak gamma-ray burst could permanently ruin our long-term potential, leaving us cowering in the dark, cold and hungry. Or maybe a bubble of true vacuum forms and gravitationally collapses Minkowski space-time as it spread out. There is, so far as I know, nothing that could be done to prevent such an event.

*
* *

Bearing all of this in mind, what ought we do? In large part this depends on what morality has to say about: *(a) the badness of extinction & (b) ideal population size.*

Moreover, there are two types of moral responses one could think appropriate in this context.

On the one hand, there is the well-worn method of unpacking these issues that's built on the back of an axiological framework. And, to be sure, this has been the bread and butter of philosophers working in the area of population ethics. Following an axiological system, the matter is reducible to 'what is good' for persons (or instead, as is often the case, in impersonal terms). Some outcome is overall better (worse) than another outcome in the set of available outcomes if it's more (less) good for persons (or the population).

This method requires getting to the bone on certain metaphysical issues, such as our modal operator (e.g., only necessary persons count towards outcome goodness)[5] and theory of goodness (e.g., the 'best things in life'[6] or the fact that we are alone in all

[3]This is a controversial topic. Some of us hold the belief that so long as humanity does not go extinct, then all development trajectories will go asymptote to the very good. This is sensible insofar as one is prepared to maintain that humankind, as it grows up, will increasingly have the power and wisdom with which to make life good. See (Bostrom, 2003a); (Bostrom, 2012a); cf. (Beckstead, 2013b). By contrast, others, such as Nick Beckstead, hold that possible outcomes in which humanity does not go extinct range continuously in their goodness. I belong to the second camp.

[4]Take the simple toy example of a totalitarian government taking control of our world after some long, terrible war which left every defeated state still standing in utter shambles. The *Supreme Leader* shuts down any and all civil protests or conflict with extreme measures. His will essentially becomes a timeless singleton—having tamed all past, present, or future opposition—which effortlessly now guides mankind's collective efforts towards achieving his personal goal of colonizing space, and putting up a flag of his face on every habitable planet. Everyone but his closest personal staff and counsel are miserable slaves whose ability to fight back has systematically been undermined (perhaps even genetically) in this possible world.

[5]See (Singer, 1976); cf. (Singer, 1999).

[6](Parfit, 1986)

of space could amplify the goodness of an outcome).[7] There is furthermore the matter of adding it all up. Our ranking of outcomes will look different on some aggregative procedures. To illustrate, if one subscribes to the Totalism package, then he will claim that our failure to prolong mankind—even if every member's life were barely worth living!—is astronomically wasteful in moral terms.[8,9]

On the other hand, one could unpack this presing moral matter with reference to 'what we owe each other'. Here once more we find that our options break down even further. The view we adopt could be held either separately from or (as part of a pluralist view) in conjunction with our above view about outcome goodness.

How the burdens of our risk-plagued world are distributed, I think, makes for an unusually strong candidate in the latter regard. *Specifically, it seems reasonble to maintain that considerations of fairness should strengthen or weaken competing claims for the adoption of some alternative.* Even if I might be as well off in expectation if my forebears took some perilous steps towards increasing their own lifetime welfare, my claim against being exposed to a risk of harm may be stronger than my forebears' claim for a chance of benefit for themselves.

It seems appropriate to consult principles of fairness as part of our moral evaluations when and insofar as there are competing claims being made by persons that cannot be measured alone by how well off these persons would be in some set of outcomes. Indeed, there are powerful arguments to the effect that without considerations of fairness, moral evaluations will struggle to account for the moral standing of individuals within the aggregative value of an outcome.[10,11] In short, perhaps outcome goodness depends (to some extent) on how proportional relations, such as 'worse off than' or 'equally well off as' with respect to the burdens of our risk-plagued world, would manifest amongst the (timeless) population.

As far as I can gather, this second approach has gone woefully underutilized in the field of population ethics. Yet, it seems to me very important that we thoroughly explore every source of moral guidance on this gloomy topic. After all, if we fail to account for all the morally relevant information, and mess things up, then there might not be a second shot at getting things right, given all of humankind could follow us to the grave.

[7] See, for e.g., (Kahane, 2014).

[8] Indeed, it is the worst possible thing we could do. For the *Astronomical Waste Argument* see (Bostrom, 2003a). (This position traces back at least as far as Henry Sidgwick (Parfit, 1984, 454).)

[9] A similar conclusion can (arguably) be reached if one is instead unsure as to which way he bends on the matter of what we ought to do. See especially (Greaves and Ord, forthcoming-*c*) for the argument that, under the Expected Moral Value model of Moral Uncertainty, our evaluations of outcome goodness will get swamped by some Critical-Threshold axiological framework in the large population limit. See (Kaczmarek and Plant) for a critical review of their arguments.

[10] (Otsuka and Voorhoeve, 2009); cf. (Crisp, 2011); (Beard, forthcoming)

[11] There might well be very many more reasons for including considerations of fairness in this gloomy conversation. Chief among them, though, I submit, is that fairness tends to survive most moral disagreements. It is about as commonly held as is the moral belief that 'improving persons' welfare is good'. That we ought to promote fairness between each other is so deeply ingrained in our regular interactions that even cute lil' monkeys have been seen to abide by its simple code (Brosnan and de Waal, 2003).

1. PRÉCIS

My dissertation takes, in a nutshell, a first baby-step towards getting traction on the second approach. More specifically, I explore what the *Competing Claims View* has to contribute to our understanding of these grave moral issues. Below you will find Fleurbaey and Voorhoeve's original formulation of the view.[12,13]

> *Competing Claims View*: we decide between alternatives by considering the comparative strength of the claims of different individuals, where
>
> (i) a claim can be made on an individual's behalf if and only if his interests are at stake; and
>
> (ii) his claim to have a given alternative chosen is stronger:
>
> > (iia) the more his interests are promoted by that alternative; and
> >
> > (iib) the worse off he is relative to others with whom his interests conflict.

*
* *

My project breaks down in two parts. The first part is by far the more demanding in terms of argumentation, and eats up the most space.

In part 1, I will argue that the best method for evaluating a set of outcomes with reference to the principles composing the Competing Claims View is the axiological framework of Averagism.[14] At first blush, this will seem dead wrong, and rather obviously so for some of my critics. After all, compare the following two vectors (this is to say, distributions of lifetime welfare in a population):

B = {1, 1, 1, 1, 1, 1, 22},

B* = {4, 4, 4, 4, 4, 4, 4}.

[12](Fleurbaey and Voorhoeve, 2012, 397)

[13]Simon Beard has since proposed a slighlty modified version—see (Beard, forthcoming)—and applied it to a moral issue which bears extreme similarity to that which I try to tackle in the thesis. Like Beard, I too will not stick to the original formulation of the Competing Claims View; however, *pace* Beard, I do not accept that the worst off person has a stronger claim insofar as he is 'more deserving' for simply being worst off. Moreover, on my version of the view, the pull of 'equal chances of a benefit' is considerably weaker, and, in fact, I proceed as if it made no distinctive contribution to our evalution of overall fairness. In its place I propose a novel (and arguably better) way to measure for fairness-weighted outcome goodness. For lack of a better idea, I call this 'burden worthwhileness'. Proponents of the 'equal chance of benefit' view include (Fleurbaey and Voorhoeve, 2012, 395); (Broome, 1990-1991); and (Otsuka, 2017). The great grandfather of the sort of toy example relied on by these authors is Diamond. I present his version of the problem below.

[14]Something that I did not have the space to properly investigate in my thesis is whether Variable Value Views, of the sort that trace back to (Hurka, 1983), could pull off the same trick (or better). My gut tells me that they would be inconsistent with our principles of fairness, given that Variable Value Views bear similarity to Totalism for short-lived populations. Still, having not had the chance to make the comparison, I cannot confidently report that this would indeed be the case. See also (Ng, 1989) and (Sider, 1991).

Some of us will have the intuition that B is less fair than B*. This is because (*inter alia*) B* suppplies an equal distribution of the benefits to their lifetime welfare. The problems do not end there. Indeed, we might find there is an issue of fairness which arises a level up from his prospects: between lotteries.[15] Consider the following options with their own associated possible outcomes in Table 1.1.

$\mathbb{L}$		$\mathbb{L}_1$	s_1	s_2
Ira	5	Ira	5	10
Eli	10	Eli	10	5

Table 1.1: Diamond's Case

Under $\mathbb{L}$, Ira is certain to get 5 and Eli is certain to get 10. By contrast, under $\mathbb{L}_1$, there is a half chance of Ira getting 5 while Eli gets 10, and a half chance of Ira getting 10 and Eli getting 5.

According to Averagism, it does not matter who gets 5, and who gets 10 under either lottery. More so, $\mathbb{L}$ and $\mathbb{L}_1$ are equally good so far as expected average lifetime welafre goes—in short, the randomly sampled person's prospects are just the same in $\mathbb{L}$ and $\mathbb{L}_1$, and therefore he remains indifferent between the two lotteries.[16] But this will stick in some of our craws, as it in fact did for Diamond.[17] He contends that $\mathbb{L}_1$ is (sometimes) fairer than $\mathbb{L}$ for the reason that the eventual distribution results from randomness, and isn't predetermined.[18,19]

While they are no storm in a teacup, I'll argue that these counter-replies are misguided. No such ingredients have to be baked into our evaluative standard in order to capture considerations of fairness in the context of how the burdens of our risk-plagued world ought to be distributed. All that is needed is the *Veil of Ignorance Principle & Conditionalism* for determining fairness from the ground-up—that is, for both: (a) individual outcomes; as well as (b) the choice between lotteries.

Part 2 of the dissertation then plays the role of determining which development policy-option (i.e., lottery), given some conservative assumptions, does best with respect to satisfying the Competing Claims View under conditions of uncertainty.

There are roughly three kinds of development policies we might adopt. We will suppose that our choice is binding; beginning with our earliest ancestors, this policy-option will be abided by for all history to come. Moreover, for the moment, let's assume there are only two states of affairs that could obtain, culminating into six possible

[15] *Clarification:* We shall follow McCarthy, Mikkola, & Thomas in distinguishing between 'prospects' and 'lotteries' as follows. We will write $\mathbb{W}$ as describing a set of welfare levels, and $\mathbb{H}$ as describing a set of histories—where a 'history' is an assignment of lifetime welfare levels to individuals (i.e., a welfare distribution) (McCarthy et al., 2016, 5). Besides lifetime welfare levels and histories *per se*, there is the probability measures over them (thereby, we assume that $\mathbb{W}$ and $\mathbb{H}$ are measurable spaces). Probability measures over $\mathbb{W}$ are referred to as *prospects*, and those over $\mathbb{H}$ are called *lotteries*.

[16] As noted before, in a previous footnote, some of us believe that the absence of a chance for Ira to benefit in $\mathbb{L}$ contributes to its unfairness.

[17] See (Diamond, 1967).

[18] This feature has often been referred to as *ex-ante equality*.

[19] cf. (Sher, 1980); (Wasserman, 1996)

outcomes when paired with our policy-options. Things could either go as planned or totally backfire.[20]

> A. *Safe-n-Slow.* If our forebears take steps to militate against the threats plaguing our world, then life may be very, very good further down in history. Yet, something might go wrong along the way, and not only will our forebears have suffered for nothing, but either (a) humanity goes extinct in the short-run or (b) civilization collapses (perhaps even several times), and though humanity recovers, life is barely worth living at several large segments of history. Here we find that our forebears have a claim against shouldering these burdens as no one may benefit from them.
>
> B. *Go-Fast.* If instead civilization ignored the threats to humanity's long-term potential, focussing on improving their own wellbeing at all times, then while humanity would go extinct sooner, those that get brought into existence would be very well off—much better off than if they instead took the occasional step to mitigate such catastrophic risks. On the other hand, humanity might survive a looming catastrophe that reared its ugly head. If so, their lives would be unbearably bad. Very many generations could suffer in this horrible plight. Here we find that our forebears have a claim for taking this option, given that it gives them chance at benefit. But there is an internal tension on this lottery which we will have to resolve (in chapter 5): while they may never come into existence, those that do and suffer the harsh conditions of a broken world have a claim against it.
>
> C. *Unambitious.* Driving the ball down middle-field, Tim Mulgan proposes the following policy-option. Consider an unambitious population who make a conscious decision to abandon any thought of colonizing the stars, and concentrate their energies on building up stable liberal institutions, accept that they will miss out on many potentially life-enhancing discoveries, and also accept that a freak catastrophe might cause extinction (or a broken world) at any moment. (But who also judge—not unreasonably—that the risk of anthropogenic catastrophe would have been much higher if they had instead pursued any more technologically ambitious policies.) In a nutshell: *humanity neither allows for cumulative man-made risk to jeopardize subsequent generations nor does it overwhelm itself with onerous burdens in an attempt to prolong human history indefinitely (or as near to eternity as one gets, given physical eschatology) under this policy-option.*

Having been armed, in part 1, with a method for ranking outcomes in terms of the competing claims of personal goodness (weighted for fairness), it is possible for us to determine the expected value of these lotteries.

[20] I lift this assumption in the actual toy model that I'll construct later on, such that, our analysis applies to real life where there are countless possible outcomes under every policy-option that will vary in their goodness. Bear in mind, all three of our policy-options can pan out a countless number of ways in the real world, given even the tiniest perturbations to a state of affairs.

Using a toy model that I programmed which simulates a large sample size of possible human histories, it's demonstrated that Safe-n-Slow does best. (Importantly, on my toy model, we were able to rule out Unambitious right off the bat, given that it's inherently less fair, all else being equal, when compared to longer-lasting populations, according to my earlier arguments. In its place I'll instead consider a hybrid of Safe-n-Slow and Go-Fast that cannot be ruled out from the armchair.)

Full disclosure: the results generated on my toy model depend on the assumptions that I plugged in. Things will look very different if you plug in something else. For instance, you might be more pessimistic than myself about the growing wisdom of humankind, decreasing astrophysical hostility in time, or our ability to acquire suitable energy during the Dark Era. But the assumptions I have made are not ad hoc (or worse)—among other things, I have taken great lengths to guarantee that what I have said is consistent with our best understanding of physical eschatology, astrobiology, and so forth. Nevertheless, though I defend my assumptions along the way, I'm acutely aware that the view from my armchair for what the universe has in store for humanity throughout all of history is obviously horrible.

At any rate, the real payoff of my doctoral project is part one. Specifically: *whatever we plug into the toy model at the end of the day, we can be confident that the output will be consistent with our principles of fairness*—or so I argue at any rate. And perhaps someone else, if not me, will someday go on to try out other combinations from the full range of assumptions available on my toy model, such that, we have a more comprehensive guide to what the Competing Claims View as such dictates.

*
* *

Here is a short prospectus.

Chapter 2 presents a pivotal bit of machinery in my argument—the *veil of ignorance*. Defined as a principle, the veil of ignorance identifies the moral point of view with the point of view of rational self-interest in the face of self-locating uncertainty.[21] This allows me to capture what I will call 'conditions of fairness'. This is because the interlocutor's 'circumscribed self-interest' forces him to be awake to every person's interests or claims, on the basis that he might end up being any one of them.

I will use John Rawls' veil of ignorance argument as my stalking horse. Bear in mind, though, that I have not set out to defend his theory of justice, and I have no plans on engaging with the voluminous literature that has since germinated. This is because, contrary to Rawls, I appeal to the Veil of Ignorance Principle[22] (hereafter abbreviated VoIP) that is axiological; this is to say, it concerns what is good for individuals.[23]

[21]Importantly, this is not the only way to define it as principle. E.g., (Harsanyi, 1953) and Vickrey (1945).

[22]So named by Teru Thomas.

[23]This being said, I make matters more difficult for myself (so as not to be accused of rigging the game). Specifically, I reject what Jeff McMahan calls 'existential benefits' (McMahan, 2013). As part of accepting the Competing Claims View, I am also commited to two items that concern our theory of prudential value: (a) Existence Noncomparativism; and (b) a Person-Affecting View. Essentially,

Specifically, I accept:

> *The VoIP for Lotteries.* Lottery $\mathbb{L}_1$ is at least as good as lottery $\mathbb{L}_2$ overall, if and only if it would be at least as good for an individual to face the prospect $\frac{1}{n}(\mathbb{L}_1(1) + \mathbb{L}_1(2) + ... + \mathbb{L}_1(n))$ rather than the prospect $\frac{1}{n}(\mathbb{L}_2(1) + \mathbb{L}_2(2) + ... + \mathbb{L}_2(n))$.[24]

On the one hand, this ought to be uncontroversial. Bear in mind, the Competing Claims View requires us to act from impartial concern for *personal (rather than impersonal) goodness*. It's very backbone is the following: *we decide between alternatives by considering the comparative strength of the claims of different individuals.*[25]

But, on the other hand, as my earlier comments made clear, some of us contest that these two lotteries would be fair in all cases (e.g., if one has equal chance of benefit, and the other doesn't). Indeed, the three principles which entail the VoIP[26] are up for grabs.[27] I'll argue in chapter 4 that they are, in fact, just right for the job.

My reason for beginning with Rawls is this. It not only provides a solid introduction to the topic of fairness and variable populations, but also facilitates my set-up of the fundamental problem that arises while attempting to secure a theory of prudential value behind the veil of ignorance. Specifically, it's not obvious how we should understand the prudential value of a lottery when he does not exist in some possible outcomes. Moreover, despite betraying Rawls—having initially set out to defend his theory of justice in this thesis!—I really do like his suggestion that the interlocutor pick the option which he would have wanted his forebears to have adopted, and this suggestion suffers from a rather fatal flaw which I'll present in the chapter.

Chapters 3 & 4 tackle the two problems I have just mentioned in reverse. The problem of prudential value is tackled by chapter 3. I adopt what Teru Thomas has dubbed *Conditionalism.*[28] According to this position, we ought to rank someone's prospects by their value conditional on his existence—so, in other words, we rule out non-existence and rescale the probabilities accordingly. Along the way, I offer a helping

outcomes are compared in terms of their being 'better for', 'worse for', or 'the same as' for individuals, and a person's non-existence has no welfare assignment. Their combination thereby renders the non-existence outcome incommensurable with any and all outcomes in which he does exist.

[24](Thomas, 2016, 130)

[25](Fleurbaey and Voorhoeve, 2012, 397)

[26]The proof for this can be found in (Thomas, 2016), as well as (McCarthy et al., 2016). A significant benefit of their argument is that their theorems are consistent with the rejection of all of the expected utility axioms, at both the individual and social levels. As they put it: "Thus expected utility is inessential to Harsanyi's approach under anonymity". This is pretty big!

[27](Thomas, 2016) argues explicitly *against fairness* in putting forward the first principle (i.e., *Posterior Anonymity*). *Contra* Thomas, I will argue that fairness has just been misunderstood in at least the kinds of variable population cases of interest to us here.

[28]Teru and I arrived at the very same result independently despite semi-regular contact throughout my doctorate—though, without a doubt, Teru did a more elegant job of the arguments. Moreover, it is now my understanding that the general idea of Conditionalism is older than that even. For example, Harsanyi considers something very much like it in correspondence (quoted in (Ng, 1983)), and more recently the view has been explored by (Voorhoeve and Fleurbaey, 2016).

hand to Rawls by describing an extant decision theory which gets him what he wants. Chapter 4 closes part one of the dissertation by first arguing (indirectly) for the VoIP, and thereby strengthening Teru's stochastic dominance argument. Next, I demonstrate that the axiological framework of Averagism describes outcome value consistently with the Competing Claims View. Chapter 4 also defines 'overpopulation' and argues that we have reasons of fairness to prevent overpopulation.

Part two begins with chapter 5 clarifying the content of policy-option (B): *Go-Fast*. You will recall, there is an internal conflict built right into this option. In order for this option to have strong enough claims, on balance, behind it (such that it doesn't lose outright on my toy model), we will understand it as composing two parts: (a) the Non-Greedy Principle of Personal Good; and (b) the Prohibition on Miserable Mere Addition. Putting them together forms what I will hereafter call, after its leading proponent, *Broome's Intuition About Neutrality* ('$\mathbb{BN}$'). I will argue that there is at least one combination of a mathematically well-behaved axiological framework and bridge principle that yields a moral theory which satisfies the normative reading of $\mathbb{BN}$. Armed with this moral theory, we're able to pull out a more precise understanding of how dangerous humanity is prepared to let things get under go-fast.

Chapter 6 runs the toy model. I rehearse some of the preceeding material, explain my assumptions, supply the algorithms used, and discuss the final results. Part two closes with chapter 7 summarizing the core findings of my research project. *Several related issues which would have distracted from my main points are critically explored in the appendix.*

Part I

The Competing Claims View

2

Fairness & 'The Veil of Ignorance'

> You can't help respecting anybody who can spell TUESDAY, even if he doesn't spell it right; but spelling isn't everything. There are days when spelling Tuesday simply doesn't count.
>
> A. A. Milne, *Winnie-the-Pooh*

This chapter introduces the basic idea of a veil of ignorance with John Rawls' Original Position (hereafter abbreviated '$\mathcal{OP}$'). Besides the fact that I really like his theory (without subscribing to it), it's simply nice to have some history about the veil of ignorance under our belts.

To repeat, Rawls' theory is very different from my own. The only real connection beyond the veil of ignorance is that we both share the goal of determining what we owe each other as a matter of fairness. You will recall, there are roughly two methods one might take in answering that question. Rawls proposes to determine what fairness requires separately from axiology—it isn't about what is 'good for persons'. By contrast, I take an axiological framework as being the initial source of outcome goodness. In short, we should always start by first evaluating how good an outcome is for persons (independent of (*inter alia*) it being better or worse for him than another outcome). Claims that could be made by a person for having some option adopted over another then further contribute to that outcome's overall goodness. Claims built atop considerations of fairness play such a role.

Though there is much to like about his theory, there are two things that'll rub some of us the wrong way. The first of these problems is largely cosmetic. The histories—this is to say, distributions—considered by Rawls do not factor in the burdens of a world plagued by catastrophic risk. His model is furthermore unable to tackle some of the most crucial issues in the field of population ethics: *badness of extinction & overpopulation.*

Unlike the first problem, it is not obvious how Rawls might fix the second problem with his toy model, the $\mathcal{OP}$. I call this the Problem of Temporal Bias. There are two horns on this beast:

> *A. There is no risk of 'digging his own grave' so to speak, given that the interlocutor's choice has no causal influence on what his forebears will have done. By leaving nothing behind for his progeny, he does better no matter what his forebears themselves left behind for him.*
>
> *B. The interlocutor will not have reason to alter the past—after all, even if history has been cruel, this might be the only outcome consistent with his existence.*

Unpacked, (A) tells us that if the interlocutor's choice cannot influence his forebears' choice, he is better off, no matter what state of affairs obtains, by leaving behind nothing for his progeny. But if his choice *does* influence his forebears' choice, then we get gored on (B)—the interlocutor ought not be bothered with which state of affairs obtains, given outcomes in which he doesn't get brought into existence are incomparable with reference to his wellbeing. The interlocutor, in short, cannot be made better or worse off if history went smoother (or rockier).

My own toy model suffers from a problem bearing extreme similarity to (B).[1] So, it is helpful to get this problem out now as part of our introduction to the veil of ignorance. This is not so with (A). Still, because I do really like Rawls' theory, and given that there is a non ad hoc solution available to him, I present the problem here so that I can, in a subsequent chapter, get rid of it on his behalf.

Below, I'll proceed by rehearsing some, but not all of his theory of justice as fairness in section 2.1. After providing the relevant background, I turn in section 2.2 to the well-worn problem addressed within Rawl's framework: *Problem of Just Savings*. His response to this problem—i.e., that the interlocutor choose as he would have wanted his forebears to have chosen—is what gives life to the Problem of Temporal Bias, and those details are unpacked in sections 2.3 and 2.4.

2.1 Background

According to Rawls, the principles of justice are those that would be chosen in a hypothetical contract. Important to bear in mind is that he writes that the $\mathcal{OP}$ is meant to be a parable, a kind of toy model to which we can turn to either remind ourselves or confirm what the demands of justice (as fairness) are.[2] To this end, the $\mathcal{OP}$ is constructed as an impartial choice situation. Because the $\mathcal{OP}$ replicates conditions of fairness among the interlocutors negotiating the terms of their contract, the principles this process delivers are just.

[1] Teru Thomas calls it the *'Risky Existential Question'*: "What should we say about the value for an individual S of a prospect in which there is a chance that S does not exist (Thomas, 2016, 103)?"

[2] (Rawls, 1971, 15, 514); (Rawls, 1993, 24); (cf. Dworkin, 1977, 150ff)

Fairness, in other words, is being baked into the procedure. But this procedure doesn't on it own entail that the chosen principles are legitimate. In order to be legitimate these principles must furthermore be *stable*: that persons, of their own accord, adhere by them in real life.[3]

In constructing this impartial choice situation, Rawls has to address the following three questions:

(a) Who is taking part in the $\mathcal{OP}$?

(b) What are the interlocutors allowed to know?

(c) How do they set about making their choice?

Rawls said very little about the hypothetical contract in relation to what we owe future people in A Theory of Justice. His thoughts on the subject are limited to §44, well after he has defended his theory of justice as fairness in another context. What we owe future persons, though not quite a storm in a teacup for Rawls, seems to have been treated as something to which his earlier arguments could be easily applied.[4]

In this short notice, we are told to adopt the "present time of entry interpretation," by which Rawls means that the contract is being formed by members of a single generation in the world's history. According to him, possible persons (or the dead for that matter) are not to be considered as taking part in this committee, nor are their interests to be represented, on the grounds that "to conceive of the original position [in this manner] is to stretch fantasy too far; the conception would cease to be a natural guide to intuition."[5]

Rawls' reply to (a) is not without its critics.[6] However, all the proposed solutions are just as bad, if not worse. I won't pursue them here. Instead, I will carry on getting the remaining ingredients on the table.

In response to (b), Rawls said that the $\mathcal{OP}$ is characterized by the *veil of ignorance*. The veil is *thick* in two respects. First, members of the assembly are to be left unaware of their personal preferences, talents, genetic endowment, status (or class), comprehensive doctrine, or where in history they hail from. This is to ensure that, in adjudicating the terms of their agreement, the contracting parties cannot exploit these (morally irrelevant) facts. Rather, they must square off on common ground. Notice, this has the effect that exactly the same pattern of reasoning will be shared by everyone behind

[3]Maintaining stability is the job of what Rawls calls *public reason*. Tackling his arguments for public reason—as well as their associated problems (e.g., the *Nightmare* (Gaus, forthcoming, 16); (Quong, 2011); (Rawls, 1999, 574); (O'Neill, 2003, note 9); (cf. Williams, 2012, 91-100)—would take me too far afield of my humble goals in the dissertation.

[4]See (Rawls, 1971, 258). He writes that 'what at first seemed a far-fetched application of the contract doctrine, his theory covers without any change in its basic idea.'

[5](Rawls, 1971, 139)

[6]See especially (Barry, 1977); (Kavka, 1975); (Schroeren, forthcoming). Some of these implications are shared by the axiology-based approach that I adopt. I do not have the space here to unpack them all. Nor would it move my argument along, given that the Competing Claims View commits me to a very particular answer. However, the interested reader may turn to (Thomas, 2016) for those details.

the veil—in other words, they behave similar to a colony of bees, a hive mentality of sorts.[7]

To be sure, it is essential to the use of Rawls' method that representatives behind the veil do not know whether they would be prepared to "bear the brunt of some chosen principle."[8] Their ignorance goes still deeper, however. Rawls contends that contracting parties in the $\mathcal{OP}$ are also unaware of everything that cannot be described in general terms. They are to be supplied only with general facts about the physical world, the evolution of life, and the contours of our moral aptitude (or, in Rawl's phrase, the laws governing human moral psychology).[9] For example, which country's astronauts flew to the moon or when the moon landing occurred (perhaps even whether we flew to the moon at all), and the number of cattle on Earth aren't up for grabs. The sole exception is that parties can be sure that, upon leaving the $\mathcal{OP}$, they'll be subject to the circumstances of justice.[10]

Second, he goes on to say that they should "discount estimates of likelihood that are not based on a knowledge of the particular facts."[11] In other words, they are supposed to ignore the principle of indifference (or what Parfit refers to as the *Equal Chance Formula*), which instructs agents to split their credence equally between two states of affairs which are evidentially symmetric.[12,13]

Regarding (c), Rawls maintains that contracting parties are neverthless *rational maximisers* of 'primary goods'—those goods the possession of which any rational agent would want more of, no matter their conception of the good life (or comprehensive doctrine). These goods are essential to free and equal persons for developing and exercising their two *moral powers*: a capacity for a sense of justice & for a conception of the good.[14] As Rawls understands them, these are those goods over which all persons will instrumentally converge. They include (*inter alia*) autonomy-preserving goods, such as rights and liberties, provisions of food and raw materials, as well as other,

[7] According to Rawls, "it is clear that since the differences among the parties are unknown to them, and everyone is equally rational, and similarly situated, each is convinced by the same arguments. Therefore, we can view the choice in the original position from the standpoint of one person selected at random. If anyone after due reflection prefers a conception of justice to another, they all do, and a unanimous agremment can be reached" (Rawls, 1971, 139). Rawls however underestimates the significance of this link between the decision making of parties in his arguments as we shall see in a later chapter.

[8] (Parfit, 1984, 392)

[9] (Rawls, 1971, 137-138)

[10] "[A] phrase under which Rawls covers such facts as that human beings are vulnerable to attack and that natural resources are limited, as are also people's powers of reasoning, memory, and attention" (Hare, 1973b, 246). (See also Rawls, 1971, 127); (cf. Barry, 1978).

[11] (Rawls, 1971, 173)

[12] (See Parfit, 2011, 350).

[13] The principle of indifferenceis not without its problems. For example, the issue of multiple partitions makes an unrestricted principle of indifference inconsistent. See especially (White, 2009) and (Greaves, 2016).

[14] The former refers to the capacity to "understand, to apply, and to act from the public conception of justice which characterizes the fair terms of cooperation. [While the] capacity for a conception of the good is the capacity to form, to revise, and rationally to pursue a conception of one's rational advantage, or good" (Rawls, 1985, 233); (see also Rawls, 1993, 19ff).

different natural resources.[15,16] Further, the parties in the $\mathcal{OP}$ are assumed to be mutually disinterested. They are strictly self-interested in buttressing their own chance at leading the good life.

At this point one might wonder how parties could make a choice that garners them the greatest share of goods (*ex ante*) in the absence of an appeal to probabilities. Well, Rawls maintains that interlocutors will be maximally risk-averse under conditions of cluelessness (which the veil supplies in abundance) and assess different principles of justice by how well off they would be in the worst case outcome. More precisely, their choice is determined by:

> *Maximin*: Maximise your expectation (in terms of primary goods) for the worst case outcome.[17]

Of course, in many aspects of real life maximin isn't a good guide for choices made under conditions of uncertainty. To borrow Hare's example:

> If, when I was a prisoner of war, a benevolent and trustworthy Japanese officer had said that he would play poker with me and, if I won enough, allow me to buy myself a ticket home through neutral territory with a safe conduct, then I should have accepted the invitation, in order to give myself a chance, however small, of freedom (the priority of liberty!) rather than forgo this chance and husband my money to buy smokes with as I languished on the Burma railway.[18]

Rawls however never intended for the maximin rule to be applied in all cases. For starters, the maximin rule is brought to bear only on the choice of principles affecting the *basic structure* of just institutions,[19] not the inner workings of home life or the peculiar dealings of prisoners of war and so on.

More importantly, Rawls claims there are three features of a situation which give "plausibility" to the maximin strategy; and "the original position has been defined so that it is a situation in which the maximin rule applies."[20] As already noted, the objective probabilities must be unknown (or, as so happens, ignored). More so, the chooser

[15](Rawls, 1971, 92)

[16]There is another primary good that Rawls may wish he had incorproated. It is 'time'. In the same way that the absence of illness is a primary good, it seems reasonable to think that it is better to have a longer rather than a shorter lifespan. I discuss this point more in the appendix with reference to Ćirković's conjecture (i.e., *Conjecture**: "[an] intelligent community tends to maximize its total number of observer-moments, *ceteris paribus*" (Ćirković, 2004a, 245)). For the moment, though, I'll note that taking stock of how long persons have to pursue the good life means our cave-dwelling ancestors were even worse off (comparatively) than we initially believed.

[17](Rawls, 1971, 153)

[18](Hare, 1973b, 250)

[19]The basic structure has a similar role to that of a constitution. It outlines the rights of persons and, importantly, what is owed to them by right. Put differently, institutions run in accordance with the basic structure are to frame persons' choice of a rational plan and incorporate the regulative element of their good.

[20](Rawls, 1971, 154-155)

must have a conception of the good life such that he is apathetic towards anything he might gain above the "minimum stipend".[21] Thirdly, some outcomes are "intolerable" or, in other words, violate the *strains of commitment*.[22] By strains of commitment Rawls is reminding us that interlocutors must choose principles they sincerely believe they will continue abiding by upon the veil of ignorance being lifted.[23] They do so by following maximin, Rawls says, because it rules out those scenarios involving unacceptable harms, where persons wouldn't be able to live with the consequences of their choice behind the veil.[24]

2.1.1 Sweeping Away Old Criticisms

Both the choice of Rawls to exclude the principle of indifference and his insistence on maximin could have been the target of my critique. However, I'm not convinced this is the best place to press him.[25] Indeed, I'll demonstrate below that Rawls may easily appeal to a *Public Order Restriction* (implied by his own approach), and brush away these potential problems.[26]

Let's start by taking stock of just how opaque these general facts are which Rawls has in mind for the deliberations of our interlocutors. Are the particular facts supposed to be referring to permanent features of the world, unalterable by the choice of our parties? If this were so, he would be quite right in maintaining that parties shouldn't have access to a fine-grained, documented history of the actual world. After all, interlocutors would then be sure that what happens in history is utterly undetermined by their choice of principle. And it's clear that in order to secure principles of justice they must imagine the history of the world as being left open for them to alter from behind the veil.

It's unlikely that these facts are all that Rawls meant to be concealed from members of the $\mathcal{OP}$. Certainly, this stance hardly seems interesting enough to have caused the kerfuffel it did in the field. More probable is that he intended *also* to obscure facts about differential progress, periods of famine in history, and all those facts surrounding our (exponentially) growing population because knowledge of these kinds of facts may undermine the impartiality of contracting parties.

If parties are granted knowledge of certain facts about life for human beings on this planet, *then* '"(even if they did not know which individuals they represent in the world

[21](Rawls, 1971, 154)

[22](Rawls, 1971, 156)

[23](Rawls, 1971, 176ff)

[24]Notice, this rationale seems to only move us towards adopting a choice-function that is maxmimally risk averse about utter calamity, not maximin (Hare, 1973b, 248-250). Following this procedure for averting catastrophic threats, their choice of principle in the $\mathcal{OP}$ more likely will bear resemblance to something like *maxipok*, which instructs our population to maximise the probability of any outcome that avoids our premature extinction (on the assumption of Existence Comparitivism) or the permanent and drastic curtailment of our population's potential (Bostrom, 2012a, 19).

[25]cf. (Harsanyi, 1975)

[26]It's my understanding that the term is owed to Carens. See (Carens, 1987, 259). See also (Rawls, 1971, 212-213).

constituted by those facts) they would be able to work out the relative frequencies of sorts of events, and thus the "objective probabilities" of occurrences [for those persons they represent]"—for instance, if apprised of the availability of (e.g.) livestock or oil, then interlocutors would be able to derive the approximate size of our population at different stages in history (among other things).[27] From this they will be able to determine that a population explosion occurred around the 19$^{\text{th}}$ century. This piece of information alone tells them they have a $\frac{1}{10}$ (or greater) chance of being alive after the 19$^{\text{th}}$ century.[28] Here, the worry from Rawls is that this will provide the grist for their tailoring principles in *unfair* ways for the benefit of persons born after this population explosion.

So, Rawls decided to keep them in the dark so to speak. On his model, instead of alerting interlocutors to the grim starting conditions for intelligent life on Earth (e.g.), we can tell them there is the possibility of progress.[29,30]

Upon reflection, however, this isn't quite right. From these objective probabilities, interlocutors are still left unsure as to how probable it is that *they* will fare well (or ill) upon leaving the rabbit hole. For *this calculation*, they must turn to the second half of what Rawls' thick veil of ignorance denies: the *principle of indifference*.[31] Indifference requires us to split our credence among possible outcomes equally when we have no reason to suppose that the probability of one outcome's obtaining is greater than that of another. For the interlocutors, this means taking the likelihood of being anyone from the total sum of persons that will exist in our population, n, as being equiprobable: $\frac{1}{n}$. From this they can then infer that their chance of being alive in or after the 19$^{\text{th}}$ century is $\geq \frac{1}{10}$. As noted, Rawls blocks the use of the principle of indifference. By taking this line, parties will no longer be able to calculate (from the objective probabilities) the expected value of adopting some policy. As Hare puts things, "Rawls does not gain anything [in this respect] by refusing to allow knowledge of these particular facts."[32]

Indeed, there are strong reasons not to refuse such knowledge. This could spell doom for our population if, for example, our parties are granted no particular details with respect to what types of crops or raw materials there are in the world. Their agreed upon principle(s) of justice would fail to specify which goods are more important for, among other things, food security.[33] To illustrate my worry, notice that things would

[27](Hare, 1973b, 246); (cf. Rawls, 1971, 168)

[28]Bear in mind, roughly 10% of all Homo sapiens who have ever lived over a period of 300,000 years are presently alive.

[29]As Tim Mulgan reminds me, Rawls also explicitly restricts his theory to the special case of societies where there is the possibility of establishing contemporary liberal democracy.

[30]Here and elsewhere my presentation (and understanding) of Rawls' view benefitted greatly from personal communication with Japa Pallikkathayil.

[31](Hare, 1973b, 247)

[32](Hare, 1973b, 247)

[33]For comparison, refer to (Tremmel, 1986) for a proposal where participants in the $\mathcal{OP}$ know the history of the real world. This includes (*inter alia*) the evolution of mankind from apes, the ubiquity of famine and pestilence throughout history, the presence of ecological disasters (both natural and anthropological), and the gradual pace of technological and scientific progress. To be fair to Rawls, he did not explicitly exclude any of this information.

go better in this world if earlier generations were to plant drought resistant crops over less stable crops. Relatedly, interlocutors should be made aware of the general curve of population growth. If previous persons were merely required, let's say, to save some ratio of their acquired goods, then larger future generations would starve.[34] In Parfit's words, "[it] is *well*-informed not *ill*-informed choices to which we can more plausibly appeal."[35,36]

However, it's not obvious that Rawls' interlocutor cannot form well-informed choices. Indeed, upon closer inspection the choice of Rawls to limit the interlocutor's knowledge to general facts about the world boils down to a matter of taste, or, alternatively, in effort of simplifying the toy model. Notice that, just as before where he allowed the interlocutor to know about the possibility of progress, we can now also alert them to the possibility of a broken world[37] obtaining, that our developmental evolution might invovlve some turbulence, and so on. So long as they are generally aware that there are threats to humanity's future on the far horizon, then interlocutors will choose policies that are sophisticated enough to (e.g.) save enough for a growing population, plant drought resistant crops over less stable crops, and so forth in order to maintain just institutions. Failure to do any of these things would be tantamount to bringing on chaos or moderate-to-radical scarcity, leading to the breakdown of public order. Indeed, Rawls claimed that liberty can be restricted for the sake of liberty *even in ideal theory*; that all liberties depend on the existence of public order.[38] Relatedly, Rawls went on to argue that if savings must be taken up to preserve the future, then doing so is required even when it comes at the expense of the worst off presently existing persons—a subject I'll return to cash out more carefully below.[39]

At any rate, this closes what has been but a tiny slice of Rawls' methodology. I have glossed over certain elements of his theory of justice (as fairness) that are, without a doubt, more complicated. Far more could have been said. Some of this will be emended as we go along. However, the preceding comments cover enough ground to get the *Just Savings Problem* going. The key features of his $\mathcal{OP}$ we must bear in mind are as follows. Membership in the $\mathcal{OP}$ is restricted to a single generation. Interlocutors make their choice using maximin from behind a veil of ignorance which affords them only the general facts about the world. Finally, the choice of policy by the interlocutors must include the public order restriction.

[34](cf. Dasgupta, forthcoming-*a*)

[35](Parfit, 2011, 351)

[36]My complaint is not new. It has been voiced before with respect to lexically-ordered general principles: "[if Rawls] limits the [persons in the original position]'s knowledge to "general facts" about the world, he is in danger of having his [persons in the original position] choose principles which may, in particular cases, result in flagrant injustice, because the facts of these cases are peculiar" (Hare, 1973a, 152).

[37]Tim Mulgan describes "[this as] a place where resources are insufficient to meet everyone's basic needs, where a chaotic climate makes life precarious, where each generation is worse-off than the last, and where our affluent way of life is no longer an option" (Mulgan, 2015, 93).

[38]See (Carens, 1987, 259). Carens dubs this the *Public Order Restriction*.

[39](Mulgan, 2011, 177)

2.2 Hullabaloo: *Just Savings Problem*

Rawls' designed the $\mathcal{OP}$ to reveal how much we ought to save for future people. And the problem of just savings is rooted in how much a population ought to consume from what they produce (as well as from the goods initially open to them).

Imagine a population at a primitive stage in our history. If this population were to consume everything, then subsequent persons (their descendants) would be left with nothing and starve. By contrast, if they consume nothing, saving everything they produce for their (possible) descendants, then they will starve (and all of humankind will follow them to the grave). Suppose instead this population consumes only what they produce, preserving for their descendants the intitial basket of goods. If this were so, then humankind would survive, but there would never be any progress made—humankind would still be stuck, foraging away on the Savanna Plains. But if they produce more than they consume, then future people will be better off.

The population we are being asked to imagine, as well as their impact on the world around them, is, of course, a huge oversimplification.[40] More realistically, there would still be leftover physical goods—at least, there would be until a point. But at the same time it's weird to think our cave-dwelling ancestors saved oil—after all, just because they couldn't access it doesn't mean they *saved* it for us. Better, I believe, is to count option value along with accessible resources. In other words, keeping existing opportunities for progress alive, as well as opening new ones, should count as part of the basket of goods saved by our ancestors. Conversely, closing possible positive trajectories counts as consuming goods.[41] At any rate, even if our forebears managed to consume everything they produced, there would be *public goods* which are either irreversible or non-consumable—for example, no one can take back curing a disease or, more colourfully, expend the invention of fire. Nevertheless, there persists a puzzle for Rawls here about how many primary goods to leave for prosperity even after factoring in these (and other) complexitites. Specifically, the problem of just savings can be reworded in terms of sacrifices made in conserving all range of goods, as well as our access to them. For ease of explication, I will continue making use of the terms 'consuming' and 'saving'.

Rawls wants to be able to say that justice as fairness requires the fourth of these options. We owe it to future persons, not just to avoid leaving them worse off than we are, but to abide by a *schedule of just savings* for posterity, such that, our children's lives will be better than ours if possible.[42]

The dilemma for Rawls is this. He agrees that savings are necessary for the benefit of mankind's long-term potential. After all, if our forebears had not saved, we would

[40] (cf. Dasgupta, forthcoming-*a*)

[41] For related philosophical work see (Sunstein, 2007, 176ff). There, he discusses the pitfalls of *irreversibility*. Similarly, Brian Barry demands "that the overall range of opportunities open to successive generations should not be narrowed" (Barry, 1978, 243). (See also Woodward, 1986, 819); (Tremmel, 1986). The skeleton of this idea has been explored in cosmology as well. For example, (Wissner-Gross and Freer, 2013) argues for the connection between intelligence and entropy maximization.

[42] (See Rawls, 1971, 285-287); (Rawls, 2001, 159).

still find ourselves in the Dark Ages. But maximin instructs parties to maximise their expectation for the worst case outcome—that is to say, they will agree on a principle where the worst off persons are given absolute (or lexical) priority. *And* Rawls committed himself to the assumption that favourable conditions would continue indefinitely, that basic needs never threaten liberties, and there would be no conflict between the liberties of present and future people.[43] Things could only get better, not worse for our population.

But if things could only get better, not worse for our population, then the earlier in history one finds himself, the worse off he will be (comparatively). Accumulation *a fortiori* involves sacrificing the interests of the worst off in human history in order to provide benefits to those that will be (*ex hypothesi*) better off.[44] Therefore, justice as fairness ought to prohibit savings; any instance of saving which might make future people better off looks like it ought to be used to improve the situation of present people, since that will be the arrangement which maximizes the position of the worst-off.

Plainly, the combination of (a) maximin with (b) the assumption of things only getting better with time is incompatible with a duty to save for a brighter future. One or the other has got to go. In trying to salvage a schedule of just savings, Rawls stuck to his guns with (b). He opted to reject maximin instead. By way of explaining this move, he argued that savings should be viewed differently from the other principles of justice. He saw it as a separate, *transitional* duty of justice, and wrote that the aim of saving is to establish just institutions (as well as to "maintain intact those just institutions that have already been established").[45]

Before describing his attempts to capture this (transitional) principle of just savings, I'll briefly outline its substance. Rawls held that savings should be compulsory up until the point where just institutions are a permanent feature of our civilization. Savings would be optional after reaching this point, given that (as Rawls believed) a just society doesn't require great abundance.[46] But he supplemented this claim with one critical qualification: future persons are never to be left worse off than their immediate ancestors.[47]

Both of his attempts to generate this principle behind the veil of ignorance share

[43](See Rawls, 1993, 297); (Mulgan, 2006, 44ff).

[44](Mulgan, 2006, 40)

[45](Rawls, 1971, 285)

[46]This would apply to 'time' as well, I imagine. I do not find purchase in this claim against abundance. My own intuition, which I believe would be widely shared, is that if some generation in (our still unfolding) history were to be presented the chance, at no or tiny cost to themselves, to sow the seeds of permanent abundance on Earth, such that, our habitat were never again subject to radical or moderate scarcity, then failing to do so is unjust. Indeed, the possibility of radical abundance may well be instantiated in our world, and rather soon in the future. The field of exploratory engineering, for instance, shows signs that atomically precise manufacturing (i.e., nanotechnology) may usher in this period of great abundance. (See Drexler, 2013). Having said that, those so inclined need not wholly reject Rawls' claim about a cap on savings. My point is compatible with saying that saving in times of foreseeable, continuing scarcity is capped but that if abundance is possible, then we owe it to future persons. And, after all, wouldn't the interlocutors deeply regret failing to have included this qualification regarding abundance in a more sophisticated schedule of savings (or cap)?

[47](Mulgan, 2011, 177)

a common feature—the *rejection of maximin.*[48] On some level, Rawls must have acknowledged that maximin could not achieve this humble goal.[49] His first attempt was to relax the strict disinterest of interlocutors in the $\mathcal{OP}$. They were to consider themselves as 'heads of families' which care to make the world a better place for two (or so) succeeding generations.[50] In other words, "[Rawls] hoped that in this spirit each generation would not display such a cavalier attitude towards posterity but rather from 'ties of sentiment' take scrupulous care for at least the adjacent generation so that, through a series of successive steps, all future generations would finally be looked after."[51]

There are two well-worn objections I'll rehearse here. First, this solution is awfully ad hoc. Isn't it strange that interlocutors have impersonal concern for future persons but not for each other? Surely, this compromises the impartiality of their choice in the $\mathcal{OP}$.[52] Second, even if we were prepared to accept this unbalanced partiality on the part of our interlocutors, this still doesn't rule out serious harms being placed on temporally-remote populations. Planting a hidden hydrogen time-bomb, rigged to detonate in 400 years, for example, is well within the parameters of acting justly (*qua* being fair to far future people).[53] More so, the suggestion from Rawls fails to account for those very many incremental harms that aggregate over the long run to constitute a grave wrong to future people (e.g., dumping toxic pollutants in the oceans or through deforestation practices while mining the Amazon Rainforest for gold).[54]

Rawls later withdrew this motivational assumption, turning instead to the following suggestion.[55]

> We say the parties are to agree to a savings principle subject to the condition that *they must want all previous generations to have followed it.* They ask themselves how much they are prepared to save should all previous generations have followed the same schedule. ... The correct principle is one the members of any generation (and so all generations) would adopt as the principle they would want preceding generations to have followed, no matter how far back in time.[56]

[48] Actually, it's not clear that he is appealing to a $\mathcal{OP}$ at all anymore. Indeed, Mulgan writes as if this is no longer the case (see (Mulgan, 2011, 176)). Others in the field make it sound like this is still taking place within the $\mathcal{OP}$ (e.g., (Schroeren, forthcoming)). Upon reflection, it's clear at any rate that you won't get a different verdict either way.

[49] In a later chapter I'll put forward my own preferred choice-function, and argue that it better captures considerations of fairness between generations. I will not however argue against the application of maximin (or close cousins of maximin). I take it that there's no need to do so, given Rawls himself shot this horse dead.

[50] (Rawls, 1971, 284ff)

[51] (Dierksmeier, 2006, 76)

[52] In personal communication Ben Colburn pointed out that even so the *impartiality of the $\mathcal{OP}$ itself* may be better promoted by relaxing the constraint on the artificial representatives it contains. Perhaps; however, I won't go into the matter seeing as how the second objection still cuts mustard.

[53] (See Mulgan, 2006, 30ff, 41).

[54] (cf. Kagan, 2011)

[55] (See Rawls, 1993, 20).

[56] (See Rawls, 2001, 160). My emphasis.

The method employed by parties in choosing a schedule of savings, then, is ranking possible schedules by deducting the cost of following a given schedule from the sum of primary goods they would have received if previous generations were to have followed it.[57]

Recall our earlier case of the primitive tribe on the Savanna. There were four options presented. Following this revamped method for determining a schedule of savings, parties behind the veil would reject complete sacrifice, as well as leaving nothing behind for future generations. But it is difficult to say what schedule they *would* agree to as part of (non)ideal theory, given the information vital to making that decision, apart from minimally being concerned with maintaining just institutions, is contingent on how the world around them happens to be.

This is however beside the point, given that Rawls' $\mathcal{OP}$ now suffers from two design flaws, one of which prevents the interlocutor from reaching this conclusion anyhow. Below, I'll begin with the cosmetic error: *his wildly misguided optimism.* The section after that introduces the two horns of the Problem of Temporal Bias on which the interlocutor now gets gored.

2.3 Misapplication of the $\mathcal{OP}$: *Rawls' Optimism*

At the time of writing his book, A Theory of Justice, the Vietnam War was raging against the backdrop of both American and Soviet Union politicians toying with each other (using the threat of nuclear retaliation) in a game of chicken. More than 3 million people were killed in the Vietnam War. Shortly before the Vietnam War took place, Nazis had ravaged Europe, along the way murdering millions in concentration camps. World War II also saw *Fat Man*, a plutonium implosion-type bomb, dropped on Nagasaki, killing more than 74,000—not to forget the catastrophe caused by dropping *Little Boy* on Hiroshima, killing almost twice as many innocent persons.

It is puzzling that Rawls' confidence was unshaken regarding the fate of humanity in the cosmos. To repeat, he assumed that favourable conditions would continue indefinitely, that basic needs never threaten liberties, and there would be no conflict between the liberties of present and future people.[58] Things could only get better, not worse for our population. As Tim Mulgan observes, "[this] optimism explains why Rawls never discussed population policy, environmental issues or the possibility that future people might be worse off."[59] Rawls saw only two problems in need of addressing: just savings and stability.[60]

But we know better than this.[61] Nevertheless, rejecting Rawls' unjustified optimism doesn't mean that we must reject his entire methodology. The Rawlsian $\mathcal{OP}$ can be reworked so that it presupposes a less rosy view of the future, by revamping its intitial

[57] (See also Rawls, 2001, 160)
[58] (See Rawls, 1993, 297); (Mulgan, 2006, 44ff).
[59] (Mulgan, 2011, 174)
[60] (See Mulgan, 2011, 178ff).
[61] And, surely, he did too—after all, why else would Rawls have introduced the Public Order Restriction?

conditions so that population policy, catastrophic risk, and so on are properly taken into account.

To this end, the obvious first step when redesigning the $\mathcal{OP}$ is to keep existential threats in mind. Moreover, to make clear behind the veil of ignorance that our world harbours many different threats that augur poorly for humanity's future. Our response to these threats must be tailored to our historical circumstances. To illustrate, the present generation finds itself in a special place in history with respect to these threats. Never before has our civilization had the ability to alter the course of history in such a permanent way. Compare slavery in the South. Of course, it was horrible and produced many still-lingering negative effects. But this appalling period of history proved to be something that our civilization could climb back from. By contrast, imagine we fail to prevent an act of bio-engineered terrorism by surviving bandits from the Aum Shinrikyo. Humanity might go extinct or find itself trapped on a grim development trajectory, culminating in something like Mulgan's broken world scenario, from which there may be no return.

Both catastrophic risk and our second big moral issue, overpopulation, could be effectively tackled with the resources already supplied by Rawls. For example, we can easily imagine overpopulation destabilizing just institutions (e.g., by producing recurring famine or war), and Rawls' proposed Public Order Restriction is supposed to prevent bad outcomes of this sort.

But there is a related issue which Rawl's approach will continue to struggle with in the context of variable populations: *how we ought to weigh a person's life against another's or against some other thing.* We may ask how the various burdens associated with catastrophic risk ought to be *compared* before we go about distributing them. There are, after all, different burdens that would have to be shouldered at different stages of history. For instance, someone will have to endure being the last of humankind (whether we go out in hellfire or in our sleep). How should this burden be set against the burden of inhabiting a broken world? Are they perhaps incomparable on Rawls' view? As things currently stand, his theory of justice seems ill-equipped for adjudicating between conflicting burdens of this sort.

That was the first of two errors which Rawls committed. Summed up:

> *1. The problem of what we owe future persons mustn't be framed in terms of how much to save for posterity, but rather with reference to which steps we ought to take towards manipulating the trajectory of humanity's long-term evolution.*

Moving forward, this brings us to Rawls' second error. Because it is a technical problem, and does not bear on his unbridled optimism about our long-term prospects, it cannot be simply removed by tweaking what the interlocutor knows behind the veil.

2.4 *Problem of Temporal Bias*

Bearing in mind that this problem has two components, let's begin with the first. (It is a problem which my own toy model does not suffer.)

Saving for the future can only leave the interlocutor worse off, given that being good to one's progeny cannot change what his own forebears have already done. The past belongs to history. Like us, our descendants are going to be stuck with whatever world we (their ancestors) leave behind for them. The parties in the $\mathcal{OP}$ are acutely aware of this. After all, the members of this committee know they are contemporaries—meaning that, their forebears are not subject to the terms of their agreement. So, if being good to their descendants won't *make it the case that* their own ancestors were good to them, then, as rational maximisers of primary goods, why shouldn't they savage the world, consuming everything and leaving behind a grim reality for possible persons that might, after all, not even exist?[62]

Importantly for our later discussion, the problem *isn't* that our progeny depends on us while we don't depend on our progeny in pursuing the good life. We might well depend on them for a number of things.[63] To illustrate, the practice of cryogenics assumes that future people will awaken us from our deep slumber. If we squander our resources, pollute the Earth, and so forth, future persons may fail to express any good will towards us.[64]

The first horn of temporal bias is instead that rational maximisers from a single generation know that, since their decision to save cannot influence whether previous persons have saved for them, there are only two possibilities. Either their ancestors saved for them or they didn't. If their forebears had saved for them, then interlocutors would maximize their share of primary goods by devouring their basket of goods. If their forebears hadn't saved for them, then interlocutors would maximize their share of primary goods by devouring their basket of goods.

[62] Bear in mind, even though the 'Heads of Family' proposal would have circumvented this problem, we have chosen to reject it for reasons discussed earlier.

[63] (Pace Mulgan, 2011, 173). As he tells it, "[we] hold their quality of life, and their very existence, in our hands, while future people can offer us nothing in return."

[64] Their intention might be to seek revenge on us for having broken their world. And one way they might express their rage is not waking up those of us that went into cryo-sleep. Other forms of revenge are also open to them—for example, branding us as monsters in whig history or worse. Some of us may, indeed, dread being remembered on the wrong side of history. (This seems, at any rate, to be in keeping with the testimonials of ex-Nazis or racists from the darkest days in Mississippi's history.) But the reasons for failing to come to our aid may have nothing to do with revenge. Take the following rather colourful example: in their search for intelligent life among the cosmos, an alien civilization comes across Earth. Imagine that they find some poorly-off humans have survived their ancestor's nuclear winter and still wander the wreckage of our broken world in search of shelter, food, and warmth. Suppose further that this intelligent alien species has the tools to reshape our world for the better. If things didn't go very smoothly for human beings—because of hate-fueled war, man-made catastrophes or catastrophes that could have been prevented (e.g., famine), and all of the other horrible, de-humanizing things we have proven capable of—these alien beings might well see no point in salvaging a species that is, so far as they can tell, inherently doomed. Or perhaps they would just lack an interest in rescuing us from our cruel fate. (These points germinated from personal communication with Nick Bostrom and Daniel Dewey.)

> *2a. Even if they are uncertain as to which of these states of affairs obtains, they know devouring their primary goods is a sure thing; therefore, they should devour away.*[65]

This may not immediately seem like a strong objection. After all, Rawls' ultimate solution to the problem of just savings is maintaining that the interlocutor ought to choose what he wants his forebears to have also chosen. More so, it doesn't matter who the man behind the curtain happens to be, given anyone that would find himself in such shoes would end up saying the same thing.

But I'll now argue that these two redeeming features are, in fact, insufficient to block the first horn of temporal bias. The interlocutor is, essentially, facing the Prisoner's Dilemma. And, as David Lewis has argued, the Prisoner's Dilemma is a Newcomb Problem.[66] Below, I reconstruct the choice situation in the $\mathcal{OP}$ to better reflect this structural analogy.

> Imagine that the interlocutor finds himself in a very tiny, very dark room. He is alone. In front of him is a table with two boxes. He can either take (a) just the right box or (b) both boxes. The left box always contains £100. The right box is initially empty. If the interlocutor takes just the right box, then it will be refilled with the previous amount plus £100 for very subsequent iteration. So, to illustrate, if he is in room 207, and no one before him has taken both boxes, then the right box contains £20,600. But if anyone takes both boxes along the way, then the right box is never refilled. It'll contain nothing in every succeeding iteration even if both boxes are taken only the one time. There is a number painted on the wall in bright yellow which corresponds to his place in this sequence. But the room is pitch black, and he cannot glean which iteration he finds himself in.

According to Lewis, the interlocutor ought to grab both boxes. This is because Lewis advocates *Causal Decision Theory*.[67] It directs agents to consider only the causal consequences of their actions. Insofar as what the interlocutor now does cannot physically alter the contents of the right box, he cannot do worse by taking both boxes. The right box either contains nothing or something. If it contains nothing, he is worse off if he leaves the left box behind. If the right box contains something, he is worse off if he leaves the left box behind. Therefore, the act of taking both boxes *dominates* taking just the right box.

However, he will exit the room and undoubtedly find that there's nothing in the right box—that is, he will walk away a moderately poor man with £100 lining his tattered pockets no matter which iteration he belongs to in the sequence. So, to be

[65] This is a straightforward demonstration of being 'under the trance of Savage's Sure-Thing Principle'. See (Savage, 1954), as well as (Pearl, 2016, 4); (cf. Stalnaker, 1968); (Lewis, 1973); (Jeffrey, 1982); (Samet, 2015).

[66] (Lewis, 1979)

[67] This theory is spelled out more precisely in the following chapter.

sure, the interlocutor will *prefer* that his forebears hadn't taken the right box, allowing its value to grow (as I'm sure Lewis himself would). This of course requires the very first interlocutor to forego any winnings, leaving the $\mathcal{OP}$ a very, very poor man. But the interlocutor after him will do as well as the first might have, and every interlocutor after that will do even better for himself. If there are very many past iterations before him, the interlocutor *would* very likely be an extremely rich man if and only if his forebears *were* to refrain from grabbing both boxes. This is exactly similar for every previous interlocutor. Everyone will reason that things would go much better for themselves if only their forebears grabbed just the right box. Thereby, one must ask how the interlocutor, knowing full-well his decision to grab both boxes is strongly correlated to what his forebears will decide, is rational to grab both boxes for himself.[68] Should he not have instead performed the act which would lead to the greatest expected value conditional on performing it? In the case so-described, taking only the right box is a strong indication of a good outcome after all.

But he can do still better (*ex ante*) if they played ball and he didn't; indeed, the best outcome from his perspective is taking both boxes conditional on all his forebears not having done the same. Furthermore, his taking both boxes cannot seal his poverty in stone. There is no causal link here; and therefore his action cannot causally influence which state of the world he does currently find himself in. So, given dominance, he will take both boxes, not being able to see the downside of aiming for the best-case scenario though he can reliably predict in advance that acts conforming to this policy will leave him a poor man at day's end.

To repeat, the interlocutor might well *prefer* that his forebears hadn't taken the right box. But this is different from what he endorses as the uniquely correct choice. His preference for his ancestors 'not ratting'[69] and leaving him well-off presupposes that

[68]There is a tendency to treat problems such as Newcomb's Puzzle as if they only arise when some other actor is a perfect predictor. This is a misconception. They can also arise when other actors merely have a partial ability to predict the agent. Indeed, we commonly find ourselves embroiled in a Prisoner's Dilemma in real life. This is because people are often pretty decent replicas of each other. "Drop a rock on my foot, and how am I likely to react? Presumably much as you would if I dropped a rock on yours" (Leslie, 1996a, 270). Digging a little deeper down, Bostrom writes,

> "[if] we imagine a space in which all possible minds can be represented, we must imagine all human minds as constituting a small and fairly tight cluster within that space. The personality differences between Hannah Arendt and Benny Hill might seem vast to us, but this is because the scale bar in our intuitive judgment is calibrated on the existing human distribution. In the wider space of all logical possibilities, these two personalities are close neighbors. In terms of neural architecture, at least, Ms. Arendt and Mr. Hill are nearly identical. Imagine their brains laying side by side in quiet repose. The differences would appear minor and you would quite readily recognize them as two of a kind; you might even be unable to tell which brain was whose. If you studied the morphology of the two brains more closely under a microscope, the impression of fundamental similarity would only be strengthened: you would then see the same lamellar organization of the cortex, made up of the same types of neuron, soaking in the same bath of neurotransmitter molecules" (Bostrom, 2012b, 71-72).

For more on this topic see especially (Lewis, 1981a); (Ahmed, 2014a); (Ahmed, 2014b).

[69]As Lewis puts it, "[the] action of "ratting" is so called because I consider it to be *rat*ional—but

they do so irrationally. Since the interlocutor endorses 'ratting' as the uniquely correct choice, he must also endorse it as uniquely rational for all previous generations.[70] This conclusion doesn't get overturned by the fact that the interlocutors are perfect replicas of each other in the $\mathcal{OP}$.

It's of course open to Rawls to simply deny either the truth of causal decision theory or its applicability in the $\mathcal{OP}$.[71] This would restore the $\mathcal{OP}$ to working order, and interlocutors would decide to save something for their progeny.

To be sure, fairness was always being baked into the procedure. But the veil of ignorance and the interlocutor's risk aversion reflected considerations of fairness. As things now stand, neither of these fairness-capturing features of the $\mathcal{OP}$ matter. Indeed, to avoid the problem of forbidden accumulation Rawls earlier chucked out maximin. Now, in order to avoid temporal bias he inadvertantly threw out the veil of ignorance as well. After all, it is superfluous, contributing precisely nothing to the decision of the interlocutor at this point. Rawls is now relying entirely on the interlocutors themselves to universalize their choice (as a categorical imperative); in short, the interlocutor is no longer strictly rational, but *reasonable* insofar as they "desire for its own sake a social world in which they, *as free and equal*, can co-operate with others on terms all can accept."[72]

However, many will see this solutions as marred by the fact that there's no obvious, let alone good reason for doing so other than to preserve Rawls' (Kantian-flavoured) intuitions about the case.[73] Indeed, this departure from the project that Rawls initially

that is controversial" (Lewis, 1979, 235).

[70]To be sure, causal decision theory isn't the only game in town. Many have criticized it for failing to 'win' in an array of colourful ingenious toy examples (e.g., the *psychopath button* (Egan, 2007); (cf. Ahmed, 2012a), *death in demascus* (Arntzenius, 2008); (cf. Ahmed, 2014a), and *newcomb's problem* (Yi, 2003)). A common objection put to the causal decision theorist is this, "why ain'cha rich?" (Lewis, 1981a)—which has been shown to backfire in the hands of the evidential decision theorist in *Arntzenius' baseball example* (Arntzenius, 2008); (cf. Ahmed and Price, 2012b)). It would take folding another doctorate into this thesis to comb through all of the alternatives (e.g., Gandalf's solution (Wedgwood, 2011)) and rigoroulsy defend a particular approach. My goal is limited to motivating a view that (a) makes you rich without 'managing the news' (Lewis, 1981b, 5) and (b) captures considerations of fairness. To this end, I take for granted that causal decision theory gets things mostly right; meaning that, causal decision theory is right to ignore spurious correlations which are irrelevant to decision making *but* it is far too strict with what counts as a non-spurious correlation. The next chapter describes more precisely the mechanics behind causal decision theory, introduces the key problems marring it, and puts forward my own preferred decision theory.

[71]Alternatively, he can argue that dominance is inapplicaple or defend some new form of ratificationism (e.g. Gustafsson, 2011). Or perhaps he could show that rational agents should take stock of more than just an act's *causal influence* on the probability of landing in certain states of the world. For instance, Rawls could weave a yarn about how the past is shaped by *quasi-causal choice*—that the interlocutor *makes it the case that* the world goes one way or another (see especially Leslie, 1996a). Alternatively, he could re-engineer the impartial choice situation such that the agent *does* have a causal impact over the decision-making of previous generations.

[72](See Rawls, 1993, 50). My emphasis. Dierksmeier summarizes things nicely: "[upon] these premises, it is evident that people who want and value social cooperation between free and equal people will accept putting themselves under such constraints which render the pursuit of their self-interest beneficial (rather than detrimental) to others" (Dierksmeier, 2006, 78).

[73]Some of us may be totally fine with this, and reject this for really being an objection to the general

set out to defend, loses all neutrality with respect to the metaphysics of the good—justice as fairness just *is* Kantianism. Having invited us into this nightmare, Rawls found himself forced to return home, to the hallowed walls of the *shining chapel of the will*, for a cure to dominance.[74] To which "[we] can only say "Amen.""[75]

However, even if we were being remarkably charitable *and* didn't press Rawls for an explanation here, he has no answer to the second horn.

> *2b. The interlocutor has no reason to prefer policies which improve the world yet are incompatible with his existence, and this will include practically every policy which deviates from the actual history of the world.*

Bear in mind, as a rational maximizer, the interlocutor prefers outcomes that are better for him.[76] If he doesn't exist in an outcome, then he cannot be well off or better off.[77] Therefore, he cannot prefer an outcome in which he doesn't get brought into existence. Changing the past jeopardizes his very existence. So, he will have no reason to abandon the policy of 'go with the flow'. The interlocutor cannot regret this decision. All his forebears will reach this conclusion as well. They too will view the default trajectory they find upon leaving the rabbit hole, where history stays the same before and after they step out from behind the veil of ignorance, as being incomparable with a more (or less) prosperous world in which they don't exist. No matter how horrible human history might prove to have been, the interlocutor will claim this is a historical record consistent with requirements of justice (as fairness). This is deeply implausible.

2.5 Concluding Remarks

As I made clear in the introduction, I used Rawls' veil of ignorance argument as a stalking horse. It not only allowed me to both introduce the basic idea behind a veil of ignorance and describe one well-known way of using it to capture considerations of fairness, but it set-up one of the key problems my own toy model must grapple with. To repeat, I have not set out to defend his theory of justice, and I have no plans on engaging with the voluminous literature that has since germinated.

---

method of reflective equilibrium. Others won't be fine with it. For example, of this general strategy on the part of Rawls, Hare writes, (and we must imagine him as frothing at the mouth) "[since] the theoretical structure is tailored at every point to fit Rawl's intuitions, it is hardly surprising that its normative consequences fit them too—if they did not, he would alter the theory" (Hare, 1973a, 147) This common charge against Rawls is given a detailed treatement in (Parfit, 2011, 346-355).

[74] I impute this clever phrase to Robert J. Steel. (cf. Rawls, 1971, 256); (Barry, 6)

[75] (Hare, 1973b, 250)

[76] *Note:* I use the terms 'population', 'outcome', 'history', 'distribution', and 'possible world' interchangeably throughout the dissertation.

[77] It is sometimes maintained that non-existence can be better for a persons that would otherwise have a miserable life not worth living. Statements of this form presuppose *Existence Comparativism*. Many—including myself—reject this position (e.g. Broome, 2004). There's no welfare level attached to non-existence, and thereby two outcomes in which the same agent only exists in one are incomparable. I have much more to say on this topic, but this must wait until chapter 5.

Still, while I have not set out to defend his theory of justice, I do have the aim of getting him out of this mess and putting both sides of the Problem of Temporal Bias to bed. The next chapter plays this role.

3

A New Decision-Procedure: *Defeating the Lil' Monster in All of Us*

> "I don't see much sense in that," said Rabbit.
> "No," said Pooh humbly, "there isn't. But there was going to be when I began it. It's just that something happened to it along the way."
>
> A. A. Milne, *Winnie-the-Pooh*

Last chapter took a wrecking ball to Rawls' ultimate solution to what we owe future people. Recall, Rawls proposed that the interlocutor pick the option which he wished his forebears had also adopted. This runs into one or the other of two horns.

The purpose of this chapter is two-fold.

First, I will put forward a new decision-procedure that is available to Rawls, and which doesn't get the interlocutor gored on either horn.[1] A ground-up renovation of the basic Rawlsian framework is simply not required.

My own toy model does not involve an interlocutor. Rather, an option is considered best if only if no other option is at least as good for persons. But we nevertheless run into the same second horn of Rawls' quagmire when we attempt to describe a theory of

[1] Importantly, there are other moves described in the Rawlsian literature—largely, Contractualist arguments—which are more closely aligned with his overall project. One departure, for instance, that my proposal requires is that future actions can be infallibly predicted. Kantians, like Rawls, will balk at this metaphysical conjecture. Among other things, it goes against the grain of what their accounts of autonomy presuppose. So, it's doubtful that Rawls would have himself wished to adopt the decision-procedure I go on to flesh out below. Still, it's worth putting it out there that the move is available to him. Plus, it is not as if we will be wasting our time—the second part of the procedure is required for my own toy model. I am grateful to Tim Mulgan and Stephan Kraemer for prompting me to produce this explanatory footnote.

prudential value in the context of variable populations. We can, luckily, help ourselves to the second half of my solution for Rawls: *Conditionalism*.

3.1 Important Point to Bear in Mind

The problem, to be sure, arises whether or not Rawls' interlocutor, while behind the veil of ignorance, knows that he will exist for sure or not.[2] Furthermore, it does not matter if the interlocutor thinks he is capable of altering history. After all, as Rawls suggests, the interlocutor makes his decision by consulting what would be best for him were his forebears to have done the same. Incomparability threatens to break the whole system down because it infects the interlocutor's counterfactual reasoning, and this happens independent of whether he knows he will exist for sure or not—'if my forebears were to ϕ, then I would be so-and-so well off'.

Bearing this in mind, I'll proceed as follows. Section 3.2. starts us off by spelling out precisely what Causal Decision Theory (hereafter abbreviated 'CDT') is committed to. You will recall, it is the application of CDT which gores the interlocutor on the first horn of temporal bias. There I go on to raise two objections which motivate the hunt for a decision-procedure that does better. Section 3.3 presents *Functional Decision Theory* (hereafter abbreviated FDT).[3] It shares most of the same features of CDT. Where it differs is in recognizing non-causal dependencies that can be crucial to rational decision-making. By applying FDT, my interlocutor avoids the first horn. But, as I'll explain in section 3.4, this then leaves him exposed to the second horn of temporal bias. In a nutshell, the problem is this: if his decision is based on an assessment of some conjunction of counterfactuals, one for every act available to him, then there will be at least some counterfactuals in which he doesn't get brought into existence. It only takes one such counterfactual to pollute the waters, paralyzing his entire analysis. I close the chapter by explaining my *partial* solution to the second horn of temporal bias. It is a partial solution because, while it solves the technical side of things, it nevertheless leaves us stuck in the mud when it comes to making any substantive claims about what we ought to do in the real world. Chapter 4 provides the rest of the solution.

3.2 Lewis' CDT

To begin, some definitions.

[2]There's one good reason for it being the latter. Derek Parfit has argued that allowing the interlocutor in the $\mathcal{OP}$ to know he will certainly exist upon exiting the rabbit hole violates the requirement of impartiality (Parfit, 1984, 392). He writes, "[the] principle we choose affects how many people exist. If we assume that we shall certianly exist whatever principle we choose, this is like assuming, when choosing a principle that would disadvantage women, that we shall certainly be men" (Parfit, 1984, 392).

[3]The theory is relatively new, building on the work of (e.g.) (Soares and Fallenstein, 2015), and has been most extensively developed in (Levinstein and Soares, 2017).

A *population* is a set of lives in a possible world. Unless otherwise stated, every mention of 'population' is to be understood as a 'timeless population'.[4] The timeless population comprises the full collection of persons across both time and space that *will* exist. It doesn't pick out some proper subset that exists at a given temporal slice (or spatial-segment). This is especially important to bear in mind because we don't want to confound the term 'population' (so understood) with every mention hereafter of a population that survives indefinitely in the cosmos. I will often slide between using the terms 'population', and 'outcome'. This is because they are interchangeable; an outcome describes the series or distribution of lifetime welfare levels of members composing a timeless population in some possible world.

I will also refer to *sub-populations* throughout the dissertation. These are proper subsets of the (timeless) population, separated from each other either by where they fall temporally or what region of space they occupy (or both). Let's say that sub-populations represented by different letters—i.e., A, B, ..., and so on—contain different lives, such that $\mathrm{A} \cap \mathrm{B} = \varnothing$. Different populations that contain some of the same persons can be constructed by the union of two (or more) subsets. The size of a sub-population is denoted $\mathcal{N}(\mathrm{A})$. Two sub-populations belong to the same population *iff* one or both have causal influence over the other sub-population. Note that the counterfactuals being run by the interlocutor involve *variable populations*. This is because the actions of his forebears would shape not only how well off persons might be, but also how many persons come into existence, as well as the composition of the population (that is, the identities of persons brought into existence). Therefore, there is no fixed population for which outcomes can be compared.

I'll follow Savage's model of decision-making in spelling out the model abstractly.[5] An agent uses his credence about which state of the world is actual to choose between possible actions that lead to better or worse outcomes. We will think of states and actions as propositions: elements of the set $\mathbb{S} = \{s_1, ..., s_n\}$ and $\mathbb{A} = \{a_1, ..., a_m\}$ respectively. When combined, an action and a state determine an outcome $o[a, s] \in \mathbb{O}$.[6] Furthermore, we will (for the moment) take for granted that the interlocutor comes equipped with (a) a (probabilistically coherent) credence function $\mathcal{P}$ that measures his subjective degrees of confidence and (b) a utility function $u : \mathbb{O} \longrightarrow \mathbb{R}$ that captures how good or bad he judges each outcome to be. Outcomes are the final ends for the interlocutor; so, $u(o_1[a_1, s_2]) = u(o_2[a_2, s_2])$ only if the interlocutor is indifferent between o_1 and o_2.

Bear in mind, CDT instructs agents to consider only the potential causal consequences of their actions. Not every correlation between act and state is relevant to decision-making according to CDT. A key ingredient in CDT is the notion of *dependency hypotheses*. Lewis says that a dependency hypothesis is "a maximally specific proposition about how the things [the interlocutor] cares about do and do not depend

[4] John Broome calls this an 'eternal population' (Broome, 2004, 18).

[5] (Savage, 1954)

[6] To take the standard example, s_1 and s_2 can describe, respectively, 'it rains' and 'it does not rain', whilst a_1 and a_2 describe 'I take my umbrella' and 'I do not take my umbrella'. o_1 then describes (e.g.) the outcome in which I took my umbrella and it rained.

causally on his present actions".[7] More precisely, it is a conjunction of counterfactuals of the form $\bigwedge_{a\in\mathbb{A}} a \mathbin{\square\!\!\rightarrow} s$. The counterfactual conditional $\mathbin{\square\!\!\rightarrow}$ is interpreted causally: if he *were* to perform a, s *would* occur.[8] So, $\mathcal{P}(a \mathbin{\square\!\!\rightarrow} s)$ is greater than, equal to, or less than $\mathcal{P}(\neg a \mathbin{\square\!\!\rightarrow} s)$ exactly in those cases where the interlocutor judges a to causally promote, be causally independent of, or causally inhibit bringing about state s.

According to CDT, these counterfactuals encode how the state of the world (as well as the interlocutor's prospects) are seen as depending on his actions. Given a probability function that is defined over a set of dependency hypotheses, CDT tells the interlocutors to maximize causal expected utility:

$$\mathcal{U}_{\mathrm{CDT}}(a) = \sum_{s\in\mathbb{S}} \mathcal{P}(a \mathbin{\square\!\!\rightarrow} s_i)u(o[a, s_i]) \tag{3.1}$$

That's about all we need in terms of the relevant background to get my critical assessment of CDT going.

*
* *

Above I said that not every correlation between act and state is relevant to decision-making according to CDT. To borrow Lewis' famous phrase, CDT doesn't embrace "'an irrational policy of managing the news".[9] This can be demonstrated by considering the following toy example.[10]

> *XOR Blackmail.* An agent has been alerted to a rumour that her house has a terrible termite infestation, which would cost her £1,000,000 in damages. She doesn't know whether this rumour is true. A greedy predictor with a strong reputation for honesty has learned whether or not it's true, and drafts a letter:
>
> > I know whether or not you have termites, and I have sent you this letter *iff* exactly one of the following is true: (*i*) the rumour is false, and you are going to pay me £1,000 upon receiving this letter; or (*ii*) the rumour is true, and you will not pay me upon receiving this letter.
>
> The predictor then predicts what the agent would do upon receiving the letter, and sends the agent the letter *iff* exactly one of (*i*) or (*ii*) is true. Thus, the claim made by the letter is true. Assume the agents receives the letter. Should she pay up?[11]

[7] (Lewis, 1981c, 313)

[8] There are a number of different ways to understand $\mathbin{\square\!\!\rightarrow}$, but there's nothing hanging on this here. See especially the essays collected in (Harper et al., 1981), as well as Judea Pearl's `do`-calculus in (Pearl, 2009) in which the agent comes equipped with both a probability function $\mathcal{P}$ and a directed graph $\mathcal{G}$.

[9] (Lewis, 1981b, 5)

[10] So far as I know, toy examples in this style made their debut in (Soares and Fallenstein, 2015).

[11] (Levinstein and Soares, 2017, 3)

Paying the blackmail is a good indication that there aren't termites. So perhaps she should pay up.

Yet, what she does now won't affect her (potential) termite problem, and will in no way improve her situation. As such, the correlation between her action and the state of the world is *spurious*.[12] This is because "paying the blackmailer when she has termites invalidates the predictive ability of the blackmailer, rather than eliminating the termites.... [if] the agent were to refuse to pay, she wouldn't change whether there's an infestation or not".[13,14] Not paying the blackmailer is the uniquely correct (rational) response.

Now, bear in mind, this line of reasoning is precisely why CDT prevented the interlocutor last chapter from grabbing only the right box. He has no causal influence on the physical contents of the box. And so, by dominance, he is rational to grab both boxes. I think this is wrong. I'll now present my two objections to CDT. To begin, I'll argue that the two cases aren't the same, and the correlation between his action (one-boxing or two-boxing) and the state of the world (the contents of the right box) isn't counterfacutally spurious. To this end, I'll provide two toy examples which buttress my conclusion. Secondly, I'll argue that although the 'why ain'cha rich?' objection may not be available to all of CDT's critics, it *does* nevertheless count against CDT.

3.2.1 *Obj. 1*: Not All Correlations Are Spurious

There is a slight but *significant* difference between *XOR Blackmail* and the two-box case. While the interlocutor has no causal influence over how many boxes his forebears took, the correlation between his action (one-boxing or two-boxing) and the state of the world (the contents of the right box) isn't at first glance counterfactually spurious in the same way. The crucial difference between the two cases can be brought out by considering another blackmail case.

[12] Put differently, it is counterfactually broken.

[13] (Levinstein and Soares, 2017, 4)

[14] Though his paper has a different aim altogether, Adam Elga's threat of torturing Dr. Evil appears to fall flat on its face for this very reason too. In his paper, the Philosophy Defense Force has sent Dr. Evil the following letter (Elga, 2004, 383):

> Dear Sir,
>
> (Forgive the impersonal nature of this communication—our purpose prevents us from addressing you by name.) We have just created a duplicate of Dr. Evil. The duplicate—call him "Dup"—is inhabiting a replica of Dr. Evil's battlestation that we have installed in our skepticism lab. At each moment Dup has experiences indistinguishable from those of Dr. Evil. For example, at this moment both Dr. Evil and Dup are readig this message.
>
> We are in control of Dup's environment. If in the next ten minutes Dup performs actions that correspond to deactivating the battlestation and surrendering, we will treat him well. Otherwise we will torture him.
>
> Best regards,
>
> The PDF

Now, this doesn't mean that Dr. Evil has no reason to be a bit panicky. But the threat won't shake him down to his rational core.

3. A NEW DECISION-PROCEDURE: *DEFEATING THE LIL' MONSTER IN ALL OF US*

> *Counterfactual Blackmail.* There is an artificially intelligent agent that plays the stock market. It's quite competent, and has amassed substantial wealth. A clever AI researcher, renowned for honesty and for the ability to predict the behavior of AI systems in simple thought experiments, acquires the source code of an artificial agent. For simplicity, assume that the researcher is a perfect predictor when in possession of an agent's source code.
>
> The researcher has developed a computer virus which will affect market operations and cause a massive crash. If the virus is used, both the researcher and the agent will lose £150 million at least. The virus is designed so that after being deployed it will remain deactivated for a day, such that the only way to prevent activation is by the agent wiring £100 million to the researcher within 24 hours. If the researcher decides to deploy the virus, they would then send a message to the agent demonstrating that the virus has been deployed and demanding £100 million.
>
> The research is very risk averse, and will only deploy the virus upon becoming quite sure that the agent will in fact pay up to deactivate it.[15]

Following CDT, the artificially intelligent agent will, once the virus is loose, fork over £100 million. This is because it recognizes that once the virus has been released the alternative is losing £150 million. So, at the time of being blackmailed, the AI would pay up. Having the ability to peer into the AI's source code, the AI researcher knows this would always be the decision of the AI. So, he writes up his letter and deploys the virus.

But if the agent were to refuse to pay, then the blackmailer would have never deployed the virus in the first place—resulting in an even better outcome.[16] Insofar as the blackmailer decides whether to deploy the virus by first accessing the artifically intellgent agent's source code, it seems that, if we can prevent him from doing so by simply lowering his credence sufficiently, then this is what we ought to do.

I'll now present another counter-example to CDT. The key difference is that this concerns decisions which are, at first blush, unstable under CDT's guidance.[17] In order to get this objection up and running I'll begin by fleshing out a few more details regarding CDT. To that end, consider:

> *Death in Damascus.* You are currently in Damascus. DEATH knocks on your door and tells you I AM COMING FOR YOU TOMORROW. You value your life at £1,000 and would like to escape DEATH. You have the option of staying in Damascus or paying £1 to flee to Aleppo. If you and DEATH are in the same city tomorrow, you die. Otherwise, you will survive.

[15] (Soares and Fallenstein, 2015, 5)

[16] (Soares and Fallenstein, 2015, 5)

[17] I should point out that the remainder of this subsection amounts to little more than an exposition of the original work performed by Levinstein and Soares—see (Levinstein and Soares, 2017, 5-8).

> Although DEATH tells you today that you will meet tomorrow, he made his prediction of whether you'll stay or flee yesterday and must stick to his prediction no matter what. Unfortunately for you, DEATH is a perfect predictor of your actions. All of this information is known to you.[18]

Plainly, what DEATH wrote down in his appointment book does not causally influence your decision. More so, your decision doesn't causally affect DEATH's decision today of where to come looking for you tomorrow. However, your decision *is* strong evidence for what the causal structure of the world is.[19] Because DEATH is a perfect predictor, a does affect the probability of $a \mathrel{\square\!\!\rightarrow} s$. It's true on CDT counterfactuals that when you stay in Damascus, you would have lived if you had fled. And it's true that when you flee, you would have lived if you had stayed.

Let S and F refer to the actions of staying in Damascus and fleeing to Aleppo, and let D and A denote the propositions that DEATH is in Damascus and that DEATH is in Aleppo. Suppose you're (initially) sure you will stay in Damascus. That is, $\mathcal{P}(\mathrm{S}) = 1$. So, you're sure that DEATH will be waiting for you in Damascus tomorrow. Since DEATH's location tomorrow is causally independent of your choice, $\mathcal{P}(\mathrm{S} \mathrel{\square\!\!\rightarrow} \mathrm{D}) = \mathcal{P}(\mathrm{F} \mathrel{\square\!\!\rightarrow} \mathrm{D}) = \mathcal{P}(\mathrm{D}) = 1$. So, fleeing to Aleppo would bring you more causal expected utility since you think you'll live past tomorrow. But once you're sure you'll flee to Aleppo, then the balances of a doomed fate tip towards DEATH being in Aleppo. In that case, staying in Damascus brings you more causal expected utility. The problem here is that cases like Death in Damascus are unstable. Either the agent is sure he'll stay put, and so goes running for the hills of Aleppo, or he ends up performing an action that violates the injunction to maximize expected causal utility.[20]

In cases of this sort, CDT broadly recommends that rational agents equilibriate their credences with respect to which action they will perform. For one, James Joyce claims that in this case the agent (a) must have non-extremal credences over his actions, given he cannot know *ex ante* what he will do; and (b) when updating his own utility calculations, he must grow more confident he'll perform attractive options but not instantly become certain he'll perform such options.[21]

Start with some intial credence function regarding what the agent believes they will do: $\mathcal{P}_0$. To put flesh on the bones, let's suppose that in this case the agent assigns $\mathcal{P}_0(\mathrm{S}) = .9$ and $c_0(\mathrm{F}) = .1$. Suppose furthermore that the utility of surviving is 1000.

[18] Death in Damascus is original to (Gibbard and Harper, 1978). This formulation is taken from (Levinstein and Soares, 2017, 5).

[19] In the same way that Johnny's decision to push the button killing all psychos is strong evidence of he himself being a psycho—see (Egan, 2007)—so he should update his credence conditional on pushing. But as soon as he updates his credence in himself being a psycho conditional on pushing, he will regret his decision to press the button; in the process violating what Arntzenius dubs *Edith Piaf's Maxim: a rational person should not be able to foresee that she will regret her decisions.* (Arntzenius, 2008, 277, 291).

[20] (Gibbard and Harper, 1978)

[21] See (Joyce, 2012). An alternative is put forward in (Arntzenius, 2008, 292ff). They bear similarity to each other insfoar as both rely primarily on the model developed by Brian Skyrms in (Skyrms, 1990). However, the differences between them are unimportant for our purposes.

If he is (fairly) sure he will stay in Damascus, then he assigns fleeing a higher expected utility than remaining: $\mathcal{U}_0(\text{S}) = 100$ and $\mathcal{U}_0(\text{F}) = 899$. After all, if he's sure he will be in Damascus, then this is where he expects DEATH to be waiting for him, scythe gleaning in the sun. Although he intitially evaluates fleeing to Aleppo as the better choice, this means that (a) he ought to raise his credence in F and lower it in S; and (b) (a) is strong evidence about DEATH's location. So, once he raises his credence in going to Aleppo, he ought to become more confident that DEATH is in Aleppo as well. So, $\mathcal{P}_0(\cdot|\mathcal{U}_0) \neq \mathcal{P}_0$, where $\mathcal{P}_0(\cdot|\mathcal{U}_0)$ refers to $\mathcal{P}_0$ conditioned on the value of $\mathcal{U}_0$. Therefore, $\mathcal{P}_0$ failed to take into account all freely available (causally relevant) information. In this case it failed to factor in the shift in his confidence regarding fleeing to Aleppo. It's clear that in cases of decision instability $\mathcal{P}_0$ should not guide his final action. After all, he can foresee regretting this act as it amounts to running straight into DEATH's waiting arms. Instead, he ought to revise his credence to $\mathcal{P}_1(\cdot) = \mathcal{P}_0(\cdot|\mathcal{U}_0)$ and reevalutate his options. He might run into the same problem, of course, in which case he ought to reiterate this process until it terminates in an equilibrium state $\mathcal{P}_e$.[22] It is only at this point that the agent has taken into account all freely available (causally relevant) information. So, $\mathcal{P}_e$ should be his credence function at the end of deliberations. Given the numbers we plugged in earlier, this process terminates in an equialibrium state $\mathcal{P}_e$ when the agent is *indifferent* between fleeing and remaining in Damascus—$\mathcal{P}_e(\text{S}) = .5005$ and $\mathcal{P}_e(\text{F}) = .4995$.

Notice that this only tells him what he rationally should believe he will do. There remains the little matter of what he ought to do. At this stage he might flip a mental coin or, following Joyce, he might be permitted to choose however he pleases. Let's suppose he flips a mental coin. If so, then "CDT agents remain in Damascus just over half the time, and they flee to Aleppo just under half the time. In each case, they end up dead."[23]

So far so good, right? Well, though this does seem like a good solution to the problem of decision instability, it means that he will sometimes end up paying money to flee to Aleppo even though he knows with absolute certainty that he's going to die there. As a matter of fact he is doomed to run straight into DEATH's waiting arms. So why is he paying £1 when he doesn't have to?[24] As if this weren't enough, things get worse if we slightly alter the toy example. This time around we will imagine that his options are (a) stay in Damascus; (b) run away to Aleppo *for free*; and (c) pay £1 for a truly indeterministic coin that lands heads with probability .5. Using this coin, he can decide whether to stay or flee. Let R refer to the act of randomizing where he will go by way of this coin. DEATH can predict whether he will buy the coin, but he can't do better than chance at predicting how the coin lands.

Surely, everyone if given the chance would buy the coin. It gives you a 50/50 shot at surviving. "By making yourself unpredictable to DEATH *even after* DEATH *has*

[22] To be precise: $\mathcal{P}_e = \mathcal{P}_{e+1} = \mathcal{P}_e(\cdot|\mathcal{U}_e)$.
[23] (Levinstein and Soares, 2017, 7)
[24] (Levinstein and Soares, 2017, 7)

made his irrevocable prediction, you end up better off."[25] Here's the rub. Levinstein and Soares have shown that CDT advises this poor soul not to pay. This poor soul is free to stay in Damascus or flee to Aleppo either way, and where he goes cannot affect (causally) where DEATH will be tomorrow. Suppose the utility of surviving is 1000. Plainly, one of these two options—staying or fleeing—has expected utility of at least 500, since staying and fleeing themselves cost nothing. Buying the coin has no causal influence on where DEATH will be tomorrow. Therefore, buying the coin costs £1, and it doesn't causally bring about a better world. Deciding where he should be tomorrow by tossing the random coin has a 50% chance of him fleeing and 50% staying, but is sure to cost £1. So,[26]

$$\begin{aligned}\mathcal{U}(R) &= \mathcal{P}(R \mathrel{\square\!\!\rightarrow} D) \cdot [.5(u(S,D) + u(F,D)) - 1] \\ &\quad + \mathcal{P}(R \mathrel{\square\!\!\rightarrow} A) \cdot [.5((u(S,A) + u(F,A)) - 1] \\ &= 499\end{aligned}$$

499 is lower than 500. Therefore, according to CDT, you ought not buy the coin. Yet, this coin is the only thing standing between you and DEATH's pervasive glare into your mind. So long as DEATH can peer into your mind, the link between his choice and your own will be counterfactually robust, and you will be doomed to run into him tomorrow.

This is the straw that breaks the camel's back, I submit. Whereas I can see someone being left unconvinced with respect to *Counterfactual Blackmail*, I struggle to imagine anyone being moved to argue *for* CDT's recommendation in this case. At most they'd bite the bullet.

Tracking the problem down to its source is simple enough. CDT is too severe in delineating non-spurious correlations. It ignores the subjunctive dependence between your choice and DEATH's decision. "*Although DEATH already made his choice before you, it seems that your choice and DEATH's choice are importantly and counterfactually linked. Because DEATH knows your mind so well, if you were to choose Aleppo (without the benefit of a truly random coin), he would have chosen it too. Buying the coin allows you to make yourself opaque to DEATH even though your choice comes after DEATH's choice.*"[27]

3.2.2 *Obj. 2*: Why Ain'cha Rich?

That was the first of the two objections to CDT that I'll make. A better-worn criticism is the 'why ain'cha rich?' line of attack. Arif Ahmed and Huw Price have formulated the argument as follows (with reference to the standard version of Newcomb's Problem):

1. The average return to one-boxing exceeds that to two-boxing (*premise*)

[25] (Levinstein and Soares, 2017, 7-8)
[26] (Levinstein and Soares, 2017, 8)
[27] (Levinstein and Soares, 2017, 8) (my emphasis)

2. Everyone can see that (1) is true (*premise*)

∴ One-boxing foreseeably does better than two boxing (*by* 1, 2)

∴ CDT is committed to the foreseeably worse option for anyone facing Newcomb's Problem (*by* 3)[28]

Given that one can foresee going home a poor man if he were to two-box, and that taking only the right box is strong evidence of a better outcome (given that all the interlocutors are perfect replicas in my two-box case), should the interlocutor not instead adhere by another leading decision theory, namely *Evidential Decision Theory* (hereafter abbreviated EDT), and perform the act which would lead to the greatest expected value conditional on performing it (that is, one-boxing)?

In response, James Joyce has argued that there is no strange mystery here; the decision to one-box is nonetheless correct as an all-things-considered strategy even if one walks away a poorer man for it on this occasion.[29]

Allow me to explain more carefully. Suppose that I abide by CDT. I am not rich because I am not the kind of person that the predictor thinks will refuse the money, unlike the one-boxer. Given that I know this of myself, and given that the predictor knows this too, it was never going to cross my mind that the box wasn't empty. The £100 was the most I was going to get no matter what I did. So, the only rational choice was for me to take it. At most, what I will have learned from going home a poor man is this: had I known in advance that I'd be in a Newcomb Problem, then I may have tried to change the type of rational agent I am (or pre-commit to one-boxing in this case) *before the boxes were filled* if I believed this might affect the predictor's prediction (and so the contents of the box). But this does not go on to provide any further reason for me to drop CDT for another decision-theory, such as EDT, which comes with its own baggage (e.g., managing the news).

Alternatively, one could respond here that cases like Newcomb's Problem reward the irrational.[30] More specifically, what Joyce's response highlights is that there is a difference between the *winning decision* and being the *winning type of agent*. Still, there is something harebrained about CDT recommending an act that returns *foreseeably* less than what alternative theories recommend. As Ahmed and Price sum things up, "[it] is no use the causalist's whining that Newcomb problems reward irrationality, or rather CDT-irrationality. If everyone knows in advance that the most productive strategy in a game is the CDT-irrational one then it is *rational* to play the CDT-irrational strategy."[31]

Of course, the dialectic doesn't stop there. Frank Arntzenius has two objections. Firstly, he asks us to imagine that there are no boxes this time around (or instead that

[28](Ahmed and Price, 2012b, 16)
[29](Joyce, 1999, 153-154)
[30]More specifically, they reward the irrational in the absence of being able to form a pre-commitment.
[31](Ahmed and Price, 2012b, 2)

the boxes are transparent). In his words, "[insane] people will see £10 in box A and £1 in box B and pick box A only. So insane people will end up richer than causal decision theorists and evidential theorists, and all hands can foresee that insanity will be rewarded. This hardly seems an argument that insane people are more rational than either of them are."[32]

My gut reaction is that this misses the point made by Ahmed and Price by miles. More so, recall the Counterfactual Blackmail case. There we saw that if a rational agent is of the type that *never* succumbs to blackmail, then able-minded predictors will always refrain from blackmailing them. In the case presently at hand, we find a similar structure: if an agent will *always* forego taking the other box, even when there's nothing that the other intelocutors can now do to prevent them from taking both (given they are isolated to their own rooms), then there's never any chance of the contents of the right box being spoiled along the way. Ultimately, then, I'm not sure I find any purchase in Arntzenius' first objection.

His second objection, however, does much better. According to Arntzenius, an exactly parallel argument works against EDT.[33] Therefore, the evidential decision theorist is hardly on good grounds to start flinging accusations around at the casual decision theorist. I won't unpack his case here. Nor will I discuss responses to that argument here.[34] This is because I have no intention of defending EDT. Rather, my goal is to show that there *is* a decision-procedure which, so far as I can tell, gets rich across most if not all cases (and which doesn't falter under the pressures of counterfactual blackmail). If this can be shown, then, while the 'why ain'cha rich?' criticism of CDT may not be available to all its critics, it will remain available to myself.

3.3 Functional Decision Theory

Functional Decision Theory ('FDT') agrees with CDT that what matters in decision-making is counterfacutal correlation based only on what *depends* on your actions. However, it also attempts to account for the kinds of counterfactual *non-causal* links that peppered the previous case studies under the umbrella of a legitimate dependency hypothesis. After unpacking FDT, I'll describe how FDT manages the five toy examples we have been working with thus far.[35] My argument in the next section is that FDT, although a massive improvement on CDT, cannot perform meaningful counterfactual analysis when some of these counterfactuals (in the relevant partition) involve the non-

[32](Arntzenius, 2008, 290)

[33]See his Red Sox vs. Yankees example at (Arntzenius, 2008, 289ff)

[34]See especially (Ahmed and Price, 2012b, 4ff).

[35]I am, of course, leaving out from consideration a catalogue of alternatives that have been proposed in the field—e.g., Wedgwood's Gandalf's Solution (or Benchmark Theory) (Wedgwood, 2011). Again, my goal isn't to rigorously compare, rehearse, or defend any particular decision theory as in all scenarios correct. I am only out to show that *this* decision theory recognizes intuitively robust counterfactual non-causal links between agents, and so avoid the first horn of temporal bias. The kinds of lessons we will learn in this section and the next suggest that another decision-procedure won't pull it off; but I'm not committed to the truth of that conjecture.

existence of the agent at hand. There I'll make the appropriate modifications so that FDT can handle this kind of decision.

A choice passage from John Leslie on the topic of *quasi-causation* will help start us off. He writes,

> Walking up to what looks like a gigantic mirror, I find myself pressing against flesh instead of glass. The universe, I conclude, must be fully symmetrical. The flesh belongs to my double—left-right reversed, but in all other respects a perfect replica. ... In crucial respects, I and my double are independent. I do not genuinely cause him to run whenever I run myself. ... None the less, by choosing to run I can *ensure* that my double runs. Without causing him to throw stones, I can *see to it* or *make certain* that he throws them. All I need to do is throw stones myself. Imagine that, seeing a bird in the other half of the universe, I want it to die. It's no use hurling a rock towards it. On reaching the place where the universe-halves joined, the rock would simply collide with the other rock which my double had hurled simultaneously. Yet what if a rock of mine kills the precisely similar bird in my half of the universe? The bird I want dead will inevitably be killed as well.[36]

Compare this to the Prisoner's Dilemma. Everyone seems to agree that the two prisoners—assuming they are sufficiently similar, and given they find themselves in the same shoes—will converge on either 'rat' or 'don't rat'. Trying to rat (kill the other guy's bird) without harming himself (killing one's own) just ends up with both serving life (two rocks colliding). The reason this is so, according to FDT, is that the two of you are running precisely the same decision-procedure. If your decision-procedure were to output 'rat', then your replica (using the same decision-procedure) would output 'rat' as well.

If you really want to kill the bird, then FDT tells us to *intervene on the algorithm, not on the action.*[37] More specifically, instead of choosing an act which maximizes expected utility where everything else is held fixed, you must see yourself as picking the best output from a decision-procedure which will also be returned by anyone executing the very same decision-procedure in the same shoes. See Figure 3.1. In Leslie's case, he ought to throw the stone at his own bird.[38] This is because where his replica ends up throwing his rock *depends* on the output of your decision function because his replica is running a copy of Leslie's decision function and has access to all the same inputs. Since the same algorithm can be multiply realized, there can be other tokens of this same (or a very similar) algorithm. And the same algorithm won't behave differently

[36](Leslie, 1996a, 270)

[37](Levinstein and Soares, 2017, 10)

[38]Obviously, the analogy to Prisoner's Dilemma starts to break apart here. But the main point is, I hope, clear. One must not rat if he wants to *make it the case that* his psychological twin does not rat as well.

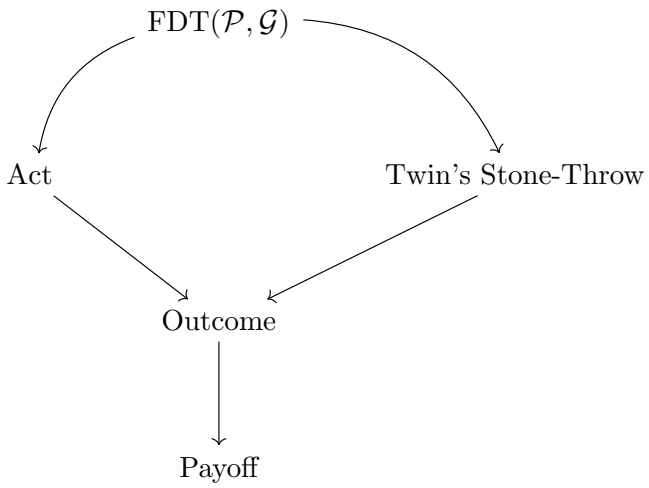


Figure 3.1: A causal graph for Leslie's Two-Birds case. If the agent wants his twin's bird dead, he should intervene on FDT($\mathcal{P}, Pg$), the node which returns the uniquely correct (rational) act, such that his twin also throws his stone at his respective bird too. Determining the best outcome by intervening at the level of the (individual's) act instead will result in stones colliding at the place where the universe-halves join.

on the same input, even if this input comes at different times or places.[39] It doesn't matter if this decision-procedure is run either very many eons in the past or in a totally alien cosmological horizon which is causally isolated. And therein is the second crucial difference between CDT and FDT. FDT agents consider only what depends on their decision, but (*pace* CDT agents) they imagine intervening not on the action directly but on the output of the algorithm that determines their action.[40]

Spelled out more precisely, FDT requires agents to have a probability and utility function along with a set of dependency hypotheses. These hypotheses are of the form

$$\bigwedge_{a \in \mathbb{A}} \text{FDT(T)} = a \mathbin{\square\!\!\rightarrow} s \tag{3.2}$$

where T is the input and s is some state of the world.[41] Put informally, an agent's dependency hypotheses are conjunctions of counterfactuals about what *would* happen if his decision algorithm on some given input were to ouput a particular action.[42] Given a probability function $\mathcal{P}$ that is defined over a set of dependency hypotheses and a

[39]This blocks the causal decision theorist's appeal to dominance. Essentially, it is a mistake to hold fixed what the other agent will do. We can only consider outcomes where there are perfectly symmetrical actions taken by the actors. So, in Prisoner's Dilemma, ratting won't dominate. This is because the only plausible outcomes are 'both rat' or 'both don't rat'. Importantly, this doesn't mean we are rejecting dominance. Much to the contrary, it simply doesn't apply in cases of this sort.

[40](Levinstein and Soares, 2017, 9)

[41](Levinstein and Soares, 2017, 9)

[42](Levinstein and Soares, 2017, 9)

directed (Pearl-style) graph encoding his views on subjunctive depedence $\mathcal{G}$, FDT tells the actor to maximize causal expected utility:

$$\mathcal{U}_{\text{FDT}}(a) = \sum_{s \in \mathbb{S}} \mathcal{P}(\text{FDT}(\mathcal{P}, \mathcal{G}) = a \,\square\!\!\rightarrow\, s) u(o[a, s]) \tag{3.3}$$

How does FDT do compared to CDT on the five toy examples?

*
* *

Well, the FDT agent will one-box in the case I described last chapter. If one-boxing is the uniquely correct output of his decision-procedure, then, given this decision-procedure is shared by every interlocutor in the $\mathcal{OP}$, the box will only contain nothing in the first iteration. In every subsequent iteration the pot will grow. Assuming there are several iterations, odds are the interlocutor will leave a very rich man. In the XOR Blackmail case, notice first that CDT won't pay the blackmail. This is because not paying dominates paying.[43] FDT agrees that you shouldn't pay up. However, the FDT agent will instead reason as follows upon receiving the letter: *the blackmailer will send this letter (a) if the house has termites if they think that the blackmail will be refused or (b) if the house doesn't have termites if they think I'll cough up. So if I refuse to pay the blackmail, then he'll never send the letter when my house has no termites. This doesn't mean I won't ever get a letter. It just means I'll only get the letter if there are termites. So really I'll only change when the letter is sent. Whilst this decision cannot solve my termite infestation problem, it does prevent me from receiving pesky blackmail letters.*[44] A small but welcome bonus. Moving along, FDT agents will never have a virus deployed on them by a blackmailer. The blackmailer in Counterfactual Blackmail can be very sure that the FDT agent will *never* pay up to deactive it. This is because if refusing to pay the blackmail is uniquely rational, then the blackmailer will reach the same conclusion when running the decision-procedure, and, bear in mind, he would only deploy the virus upon becoming quite sure that the agent will pay up to deactive it. Therefore, he will never deploy the virus.

This leaves Death in Damascus.[45] In Death to Damascus the agent is doomed. DEATH is running the exact same algorithm; so no matter the action (S or F), this will be DEATH's location (D or A). But unlike the CDT agent, the FDT agent, after settling his nerves with something strong, will decide to stay in Damascus. He will die there, but he would have died either way and at least in Damascus he gets to keep the

[43] Paying off the blackmailer cannot affect my (non)termite problem. If there are termites and I were to pay, I lose £1,001,000. If there aren't, then I'd be out £1,000. If I refuse to pay and there are termites, I would lose £1,000,000. If there aren't, I'd lose nothing.

[44] (Levinstein and Soares, 2017, 11)

[45] Several more vignettes of Death in Damascus (among other cases) are developed starting at (Levinstein and Soares, 2017, 12ff). Whilst I do not have the space to explore them here, the Asteroids in Aleppo case is particularly interesting as DEATH tries to use his knowledge of an agent's behaviour to cause bad outcomes. As Levinstein and Soares (correctly) demonstrate, the FDT agent is safe from DEATH's meddling ways here.

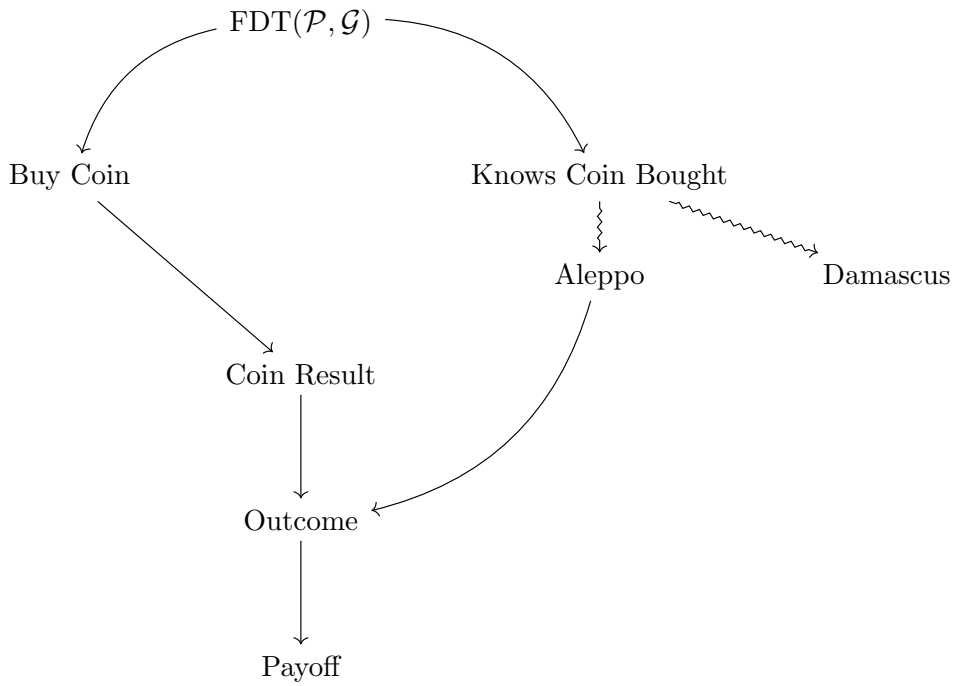


Figure 3.2: A causal graph for defying DEATH. The squiggly line describes DEATH's uncertainty beyond knowing that the coin has been purchased. We suppose that DEATH randomly picks Aleppo.

£1. Therefore, outputting S results in a better outcome. In the altered case where he has the option of buying a truly random coin, the FDT agent definitely buys the coin. In this case he does so because the best way to intervene on the algorithm is to place a wedge between himself and DEATH. See Figure 3.2. This obscures DEATH's ability to reliably predict where this soul will be tomorrow. DEATH will have at most a 50/50 shot at successfully harvesting his soul.

In both versions of Death to Damascus and elsewhere we have seen that the FDT agent either outperforms the CDT agent or does at least as well (but often for better reasons). The upper-hand goes to FDT precisely because it recognizes non-causal dependencies that can be crucial to rational decision-making. When others have access to your decision-procedure, their output is at least partially determined by your action insofar as two rational agents cannot return different (uniquely correct) answers given the same input on the same algorithm. Finally, FDT isn't vulnerable to the 'why ain'cha rich?' objection. He *is* almost certainly rich.

3.4 Problem: *Spooky Counterfactuals*

Using FDT, the interlocutor doesn't have to view his decision as being capable of altering history. 'Whatever will be will be' is fine so far as he's concerned. If we suppose this is true, then his choice cannot change his actual circumstances or jeopardize his existence. He isn't, as such, calculating how well off he *will* be if he picked some policy. Rather, following Rawls, he could be picking a policy that he wished his forebears *would* have abided by in the actual world. As an FDT agent, he is, of course, committed then to the same policy—to repeat the slogan: *intervene on the algorithm, not the action.*

Here we run up against the second horn of temporal bias. *The interlocutor will not prefer policies which are incompatible with his existence.* There are two different ways the interlocutor running FDT might find himself gored on this horn. Both however boil down to the fact that the conjunction of counterfactuals, one for every act available to the interlocutor, include some counterfactuals in which he doesn't exist.[46] In both cases these kinds of counterfactuals will pollute the waters, corrupting his decision-procedure in such a way that returns utter nonsense. This is because I commit us to the assumption that the interlocutor cannot be better off existing happily than not existing at all (or worse off existing miserably than not existing at all).[47] *His non-existence is incomparable.*[48]

The first way he might get gored is if some batch of possible worlds corresponding to some policy-choice are totally inconsistent with his own existence. He cannot compare this policy to other policies in terms of outcome goodness. This is because these associated possible worlds are neither better for, worse for, nor equally good for him than some other outcome. These outcomes are incomparable. And, therefore, the interlocutor cannot rank this policy. Upon reflection, however, there's no policy-

[46]There is no *Grandfather's Paradox* at work here. That problem arises only if the agent alone had the causal power to undermine his own existence, and in the process making it so that he never could have performed the action. The toy example is travelling back in time and killing one's own grandfather. In the case at hand the agent doesn't have this kind of causal power. Whatever he would have submitted to the algorithm will also be that which his exact replicas would have submitted. Yes, the decisions of his forebears depend on what he decides in an important and counterfactually robust sense. But the correct way to view this relation is that whatever gets fed into the algorithm is what any rational person in the exact same shoes would have input. So, it's perfectly coherent to say that 'if he were to input x as the uniquely correct rational choice, then this is what his forebears (as equally-situated rational actors) would input; so, he won't ever get the chance since he'll never be brought into existence'. So really this is at most a strange case of a counter-possible. (On the general subject of counter-possibles, Levinstein and Soares acknowledge that FDT is not a fully worked out theory just yet (Levinstein and Soares, 2017, 9-10); (cf. Bjerring, 2013).)

[47]This thesis is commonly referred to as *Existence Noncomparativism* in the field of population ethics. The rejection of this eminently plausible thesis requires assigning some numerical value to non-existence. This value could be positive, negative, or zero. Most find this position deeply weird, objectionable. See especially (Broome, 1999, 168).

[48]And if the interlocutor doesn't exist in a possible world, then we denote this with an arbitrary non-numerical value, Ω. This the de facto standard in the field post-(Broome, 2004, 25). E.g., imagine two outcomes, A = {2, 4, 2, 2} and B = {2, Ω, 4, 4}. In B the second person in this series doesn't exist, while the third and fourth persons are better off. This is an illustration of *malign addition*: the addition of a person reduces the welfare of some members of the population.

choice—including going extinct after a single generation—which is totally inconsistent with his existence. After all, he might belong to the beginning of human history, and this generation overlaps every possible history of the world.

The real trouble begins once we recognize that we must attach some probability of him not being brought into existence to every available policy-choice. After all, the interlocutor is considering counterfactuals that cover all of history. A small change here or there by his forebears is, at first blush, sufficient to block him from being brought into existence in at least one possible world. If this is so, then there's no coherent way for him to evaluate how well off he would be in expectation if this policy were followed by his forebears. Imagine that he has some (non-negligible) credence in not existing if policy x were followed. Given that his non-existence has no value, not zero value, the product of it with his credence in this counterfactual is going to be an undefined number.[49] He then must repeat this process for every other counterfactual such that he has a well-defined prospect for this policy-option, $\mathbb{L}$. Even if the rest of the counterfactuals produce well-defined numbers, the overall sum is meaningless.

Either way we cash it out, then, the second horn of temporal bias therefore spells doom for the interlocutor's application of FDT.

3.4.1 Solution

As it turns out, there's an elegant solution to the problem. Indeed, it has been inadvertently suggested by David Lewis himself. Lewis describes $\mathbb{S} = \{s_1, ..., s_n\}$ as a *rich partition* of outcomes that is fine enough to capture the differences that the interlocutor cares about.[50] For example, in the $\mathcal{OP}$ if the interlocutor only cares about winning the Queen's shilling, and not about winning for its own sake (e.g.), then $\mathbb{S} = \{s_1, ..., s_n\}$ is a partition descrbing how much he wins. A dependency hypothesis then describes a pattern of dependence of $\mathbb{S} = \{s_1, ..., s_n\}$ on $\mathbb{A} = \{a_1, ..., a_n\}$, with one conjunct for every combination of s_i and a_i.

My suggestion for Rawls is this. Suppose that *the interlocutor cares only about how well off he is conditional on existing.*[51] I plan to capture this concern of his in two parts. Firstly, the partition will describe how much more than nothing he would have in this outcome. If he doesn't exist, then he has zero more than nothing. So, we write zero. If he does exist, then he has some welfare more than nothing. So, we write this welfare level down. In this way, Ω will no longer pollute his counterfactual analysis.

But someone would be right to criticize this for being a sneaky work-around to existence noncomparativism. My response is that it would be if not for the second part. In order to capture the conditional, we will have the interlocutor assign zero credence to those outcomes in which he doesn't exist, and renormalize the probability of the remaining outcomes. In a nutshell, Rawls' interlocutor is overly-confident about his existence being preserved no matter which path his forebears would have taken.

[49] (Herstein, 2013)

[50] (Lewis, 1981c, 317)

[51] You will recall, this is a position, *Conditionalism*, which Teru Thomas independently also arrived at. There are some small differences between our accounts. See (Thomas, 2016).

Suppose that if his forebears 'ratted', then there are three outcomes he might find himself in: he might not exist; he might be miserable; or he might be moderately well off. His credence is initially split equally among them. If they instead 'don't rat', then there are only two possible outcomes: miserable; or moderately well off. Again, his credence is equally split here. Whether ratting is better than not ratting is determined by first assigning zero to the outcome in which he doesn't exist, and normalizing. Then he performs an expected value calculation. Ratting will have the same expected value as not ratting; meaning that, the outcome in which he doesn't exist does not alter his evaluation of the outcomes in which he is either miserable or moderately well off. So, in this way, the interlocutor doesn't attach any positive or negative weight to the outcome in which he doesn't exist, given his forebears choose to 'rat'. Thereby, we preserve the incomparability of non-existence.

Imagine the interlocutor is $p_{442} \in \mathcal{L}$ in the real world. Because he cannot be anyone but p_{442}, $\mathbb{S} = \{s_1, ..., s_n\}$ is a partition describing how much more than nothing p_{442} has. Following the above recipe, every world in which p_{442} doesn't get brought into existence is assigned zero and dismissed. The interlocutor then judges how good or bad each outcome is by way of his utility function $u : \mathbb{O} \longrightarrow \mathbb{R}$. Since the interlocutor's comprehensive doctrine is opaque in the $\mathcal{OP}$, his goal is to maximize his shot at obtaining the good life no matter how his comprehensive doctrine gets coloured in.[52] To that end, he endorses the option which has the best prospects with reference to p_{442}'s share of primary goods. But since he does not know if he in fact is p_{442} in a given outcome, only that he certainly exists in this outcome, the interlocutor's evaluation of how well off the i^{th} individual does in $\mathbb{L}_n$ (written $\mathbb{L}_n(i)$) becomes the same as an impartial evaluation of the average lifetime welfare in $\mathbb{L}_1$. If we assume that there is no upper-bound on how well off he would be by acquiring even more goods, which Rawls could deny only at the cost of complicating the math, then we arrive at:

> *The VoIP for Lotteries.* Lottery $\mathbb{L}_1$ is at least as good as lottery $\mathbb{L}_2$ overall, if and only if it would be at least as good for an individual to face the prospect $\frac{1}{n}(\mathbb{L}_1(1) + \mathbb{L}_1(2) + ... + \mathbb{L}_1(n))$ rather than the prospect $\frac{1}{n}(\mathbb{L}_2(1) + \mathbb{L}_2(2) + ... + \mathbb{L}_2(n))$.[53]

3.4.2 An Alternative

There is an alternative to the vanilla flavoured Conditionalism that I (as well as Teru) adopt. As Teru points out in his own dissertation, "[the] alternative is to rank prospects by their value conditional on existence *and then wieghted* by the probability of existence".[54]

This is not obviously insensible. It allows us to retain our position that non-existence is incomparable to existence while also relfecting an arguably relevant difference between a lottery A which is sure to provide 10 welfare and the lottery B which has a

[52] Well, that's a bit too strong. Rawls imposes the restriction of *Reasonable Pluralism* on his $\mathcal{OP}$.
[53] (Thomas, 2016, 130)
[54] (Thomas, 2016, 146)

.5 chance of providing 10 welfare and a .5 chance of non-existence. On this kind of case there is a rather large substantive difference between the individual's prospects on Conditionalism and Weighted Conditionalism. Specifically, the latter describes the value of B as 5 while the former says it is 10. Thereby, we are only indifferent between the two lotteries under Conditionalism.

But this position is surely mistaken for reasons that Teru Thomas fleshes out. Consider the following two lotteries:

$\mathbb{L}_1$	Heads	Tails	$\mathbb{L}_2$	Heads	Tails
Ira	10	Ω	Ira	6	6

Table 3.1: Teru's Case

As Teru argues, the proponent of Weighted Conditionalism is committed to $\mathbb{L}_2$ being better than $\mathbb{L}_1$, but, importantly, has no grounds for such a ranking in expected utility theory.

> Observe that the outcome on Heads is much worse for [Ira] under [$\mathbb{L}_2$] than under [$\mathbb{L}_1$]. That seems to be one respect in which [$\mathbb{L}_2$] is worse than [$\mathbb{L}_1$]. If the outcome on Tails were much *better* for [Ira] under [$\mathbb{L}_2$] than under [$\mathbb{L}_1$], that would be a countervailing respect in which [$\mathbb{L}_2$] would be better than [$\mathbb{L}_1$]. So we might reason: yes, on Heads, [$\mathbb{L}_1$] is better for [Ira], but, on Tails, [$\mathbb{L}_2$] is better to an even greater degree; thus [$\mathbb{L}_2$] is better overall. The weighted conditionalist *cannot* argue this way, because he maintains that 6 is not better than Ω. But then it is simply not clear to me *why* [$\mathbb{L}_2$] is better for [Ira] than [$\mathbb{L}_1$].[55]

This seems to me obviously right. We will stick to the vanilla flavour of Conditionalism.

3.5 Problem: *Double-Header*

For the moment, let's pretend that Rawls' interlocutor exists in every outcome.

A problem which immediately arises is that, even if the interlocutor can be sure he exists (or reasons as if he will surely exist), he nonetheless has utterly no idea *what his number is* or *what his world is like*. Therefore:

> (a) the interlocutor cannot define the relevant partition $\mathbb{S} = \{s_1, ..., s_n\}$; and
>
> (b) he doesn't know how to split his credence over the (circumscribed) conjunction of counterfactuals.

In essence, even though counterfactuals in which he doesn't exist no longer muddy the waters, we still find that the interlocutor cannot get down to the business of crunching the numbers.

[55](Thomas, 2016, 147-148)

3. A NEW DECISION-PROCEDURE: *DEFEATING THE LIL' MONSTER IN ALL OF US*

Let's start by getting a handle on the second head of the problem (b). Suppose for the moment that the interlocutor is p_{442}, and that he has only two acts to choose from: 'rat' or 'don't rat'. Though we may have some idea, sitting in our armchairs, of how 'ratting' will influence the world around him, there are infinitely many outcomes possible in which p_{442} gets brought into existence, given both that:

> (*i*) the interlocutor doesn't know (*inter alia*) what the initial conditions are for the earliest cave-dwellers, how severe the problem of catastrophic risk is, how hard it will be to solve that problem, the long-term potential of mankind, or how difficult it will be to improve the world around them; and
>
> (*ii*) there are an unknown number of random or causally indepedent factors which will also influence the population's trajectory (e.g., gamma-ray burst, the formation of strange matter, the freak mutation of a crop-killing bacteria).

For comparison, imagine dropping the same feather some number of times from a height of fifteen hundred feet. There's only so much that the feather-dropper can control for; the rest is terrifyingly out of his hands—e.g., a freak gust. He is almost certain never to hit the same spot on the ground twice no matter how many times he attempts it. Similarly, the interlocutor's forbears' act of ratting can give way to infinitely many possible worlds in which he is left in riches or ruin (and every shade of lipstick in between).

Even if the interlocutor knows that he is p_{442}, it's wildly unclear what may be waiting for him, conditional on 'ratting', upon lifting the veil. Again, it's not just that the interlocutor is blind to what the real world is actually like, there's also countless perturbations which both will influence his population's development trajectory and are out of his hands. Blinding the interlocutor to the knowledge that he is (e.g.) p_{442} only adds to this mess.

This being said, let's suppose that there were some non-arbitrary way to split his credence among the partition $\mathbb{S} = \{s_1, ..., s_n\}$ relevant to some p. If that were so, then Rawls' interlocutor doesn't need to define the relevant partition for a unique p (solving the problem of (a)). Instead, he can crunch the numbers for p_1, determine what the expected value of ratting is, and compare that to the expected value of not ratting. He must then repeat this process for every $p \in \mathcal{L}$ which might get brought into existence at some point in human history. *If* the interlocutor has no reason to think he is more likely to be p_{442} than $p_{444,444,442}$, he ought to split his credence between them. And if that's so, then he should just take the average expected value of ratting (and again for not ratting). He should then pick the act with the highest average lifetime welfare (in expectation).

But where he finds himself in history *does* depend on whether he rats or doesn't rat to some extent. It might be that ratting severely reduces the chances of humanity surviving beyond the Iron Age. As such, he shouldn't be indifferent when splitting his credence—after all, he *is* more likely to be p_{442} than $p_{444,444,442}$ if most worlds end before $p_{400,000}$ gets brought into existence.

So, in total there are three issues to resolve if Rawls were to adopt this method: (*i*) ratting produces an infinite number of varying worlds; (*ii*) there is no known non-arbitrary method for determining an individual i's relevant partition, $\mathbb{S} = \{s_1, ..., s_n\}$; and (*iii*) the act of ratting should swing the interlocutor's credence in the world being so-and-so by some unknown amount. So long as these three problems continue to plague the interlocutor behind the veil of ignorance, then defining his prospects under some lottery $\mathbb{L}_n$ remains untenable (unless we are content with wild speculation).

Frankly, I can see no progress being made on tackling this problem from the comforts of our armchairs.

3.5.1 It's a Problem for Measuring Competing Claims Too

While my own toy model throws out the interlocutor altogether, ranking outcomes simply by how good they are for persons conditional on their existence, it nonetheless faces the same basic issue. Bear in mind, the Competing Claims View states that:

> *Competing Claims View*: we decide between alternatives by considering the comparative strength of the claims of different individuals, where
>
> (i) a claim can be made on an individual's behalf if and only if his interests are at stake; and
>
> (ii) his claim to have a given alternative chosen is stronger:
>
> > (iia) the more his interests are promoted by that alternative; and
> >
> > (iib) the worse off he is relative to others with whom his interests conflict.

The same problem Rawls countenances gets going for my own account because (iia) requires comparing how much better or worse off an individual i's prospects are under a given lottery. And to do this we need to know how *a specific individual's* prospects—bear in mind, his prospects are the collection of probability-weighted welfare distributions under some lottery—are defined for every lottery, and then comparing the strength of every individual $i \in \mathcal{L}$'s claim against each other.

Here is one way we might step around the problem.

We would begin by noting that for every possible development trajectory, there is at least one possible world in which we can describe a lifetime welfare level for $i \in \mathcal{L}$. This, though, runs into yet another metaphysical problem. As David Lewis has argued, the very same person cannot exist in more than one possible world. At most he has a counterpart in another world.[56] Therefore, it is false that, for example, i is better or worse off in any world other than the actual one. But this is little more than a storm in a teacup. In Broome's words, "[as] Lewis agrees, if there is a metaphysical problem, it is about how [the statement 'David Lewis would have existed if the Finnish War never took place'] is true, not about whether it is true. If Lewis is right, counterparthood

[56](Lewis, 1981c, 198-202)

can substitute for identity".[57] It's, at any rate, a helpful yarn that makes sense of the personal betterness relation that we commonly understand to hold between different histories in population ethics.

If we are prepared to take this on, then one could define the individual shortfall of $i \in \mathcal{L}$ in world B as the comparative loss in lifetime welfare level in world B compared to the world X in which he does best. From there all that is needed is to add up every individual's shortfall that exists in world B. This sum then contributes towards overall value above and beyond how good world B is for persons.

However, this only seems manageable in decisions made under conditions of certainty. If we are deciding between lotteries, then we must determine the *expected shortfall* of every $i \in \mathcal{L}$; we cannot simply take the worst and best case under these lotteries as sufficiently instructive. Consider the following two lotteries:

$\mathbb{L}$	s_1	s_2	s_3	$\mathbb{L}_1$	s_1	s_2	s_3
Ira	6	6	6	Ira	2	2	14

Table 3.2: Expected Shortfall

Suppose, for illustration, that we simply calculated the shortfall of $\mathbb{L}$ by taking the outcome in which Ira does best (i.e., $14 = s_3 \subset \mathbb{L}_1$) and subtracting the worst case under $\mathbb{L}$ (i.e., $s_1 = 6$). This tells us that the shortfall of $\mathbb{L}$ is 8. Meanwhile, the shortfall of $\mathbb{L}_1$ is $6 - 2 = 4$. But is $\mathbb{L}$ twice as bad for i as is $\mathbb{L}_1$? It is not. $\mathbb{L}_1$ is in one respect worse if we accept the general notion of priority.[58] Morever, there is another respect in which $\mathbb{L}$ cannot be worse than $\mathbb{L}_1$: *Ira's prospects are exactly the same in both.*

Taking the expected shortfall of a lottery requires knowing the average lifetime welfare of i, given this prospect. Only then do we compare it to his average lifetime welfare level in the lottery in which he does best. But this requires doing what I above conceded seems impossible in the absence of a crystal-ball in my toy model (in part 2). Even if we don't weight a prospect by the probability of i's existence, there's no way for us to know the content of the infinite series of lifetime welfare levels describing his welfare in every outcome in which he exists, given some lottery. So, we cannot determine his average lifetime welfare level (conditional on his existence).

At any rate, this grist for the windwmill is simply absent in the VoIP (which only concerns itself with how good every lottery is for the randomly-sampled individual), and it is not obvious how one should carry on trying to incorporate it in our analysis now, especially given the above problem.[59]

[57](Broome, 2004, 15)

[58]Prioritarians tend to prefer $\mathbb{L}$ over $\mathbb{L}_1$ because they think the weflare difference between 2 and 6 is more important to overall value than the welfare difference between 6 and 14.

[59]Importantly, I do not mean to suggest that we cannot perform a rigorous evaluation of simple toy examples where some small number of individual's prospects are handcrafted from scratch. We can. But I worry there is little to learn from such toy examples above and beyond whether or not our theory is intuitively plausible. Certainly, they will not help us resolve the moral quandary of whether to develop fast or slow.

3.6 Closing Remarks

The chapter introduced (a) FDT (so as to extend an olive branch towards Rawls) and (b) Conditionalism. I only need the second component, (b), for my own toy model. It behooves me to now explain why pairing this up with the VoIP results in evaluative considerations that correctly describe the value of humankind's development options with respect to the Competing Claims View. Specifically, as we have just seen, it looks like we will not be able to accurately measure for the strength of competing claims made by individuals. This is the subject of the next chapter.

3. A NEW DECISION-PROCEDURE: *DEFEATING THE LIL' MONSTER IN ALL OF US*

4

Averagism as Proxy

> "This ... whatever-it-was ... has now been joined by another ... whatever-it-is ... and they are now proceeding in company. Would you mind coming with me, Piglet, in case they turn out to be Hostile Animals?"
>
> A. A. Milne, *Winnie-the-Pooh*

In this chapter, I'll argue that the Competing Claims View should also factor in the *worthwhileness* of various distributions of harms and benefits. While other considerations (e.g., equal chance of benefit) strengthen a person's claim for some option, worthwhileness is a countervailing force that weakens his claim.

Upon factoring this additional consideration into the Competing Claims View, we find that its evaluations of overall value in large population cases just about totally reduces back down to evaluations of what is good for persons. Therefore, the Competing Claims View generates approximately the same *cardinal ranking* of lotteries as a form of Averagism resulting from the combination of the VoIP and Conditionalism.

4.1 The VoIP

This section presents Teru Thomas' three principles which, his arguments show, are jointly equivalent to the Veil of Ignorance Principle ('VoIP'). Bear in mind, the VoIP holds that:

> *The VoIP for Lotteries.* Lottery $\mathbb{L}_1$ is at least as good as lottery $\mathbb{L}_2$ overall, if and only if it would be at least as good for an individual to face the prospect $\frac{1}{n}(\mathbb{L}_1(1) + \mathbb{L}_1(2) + ... + \mathbb{L}_1(n))$ rather than the prospect $\frac{1}{n}(\mathbb{L}_2(1) + \mathbb{L}_2(2) + ... + \mathbb{L}_2(n))$.[1]

This is best understood as saying that our evalutions of a lottery's overall value are equivalent to our evalutions of the prospect an individual countenances who is uncertain

[1] (Thomas, 2016, 130)

of where he falls in this history (so, has equal chances of subsequently facing $\mathbb{L}_1$ through $\mathbb{L}_n$).[2]

Viewed this way, the VoIP supplies a "different way of thinking on which overall as opposed to individual evaluations can still be 'personal'" as compared to simply what is good for one individual.[3] Whatever else some of us might think about this kind of approach, it does undeniably line up well with the Competing Claims View, at least at first blush.

Below are our three principles, as well as four diagrams to aid in your reading them.[4]

> *Principle 1.* If the same anonymised distributions get the same chances in $\mathbb{L}$ as in $\mathbb{L}_1$, then $\mathbb{L}$ is just as good as $\mathbb{L}_1$.
>
> *Principle 2.* If each person faces the same prospect in $\mathbb{L}_1$ as in $\mathbb{L}_2$, then $\mathbb{L}_1$ is just as good as $\mathbb{L}_2$.
>
> *Principle 3.* In cases of perfect unanimity, $\mathbb{L}_2$ is at least as good as $\mathbb{M}_2$ if and only if it is at least as good for each (hence every) individual.

$\mathbb{L}$	
Ann	x
Bob	y

$\mathbb{L}_1$	s_1	s_2
Ann	x	y
Bob	y	x

$\mathbb{L}_2$	s_1	s_2
Ann	x	y
Bob	x	y

$\mathbb{M}_2$	s_1	s_2
Ann	a	b
Bob	a	b

Although I omit the details here, the general proof that these three principles (along with some domain conditions—i.e., conditions to the effect that all relevant lotteries exist—and technical assumptions 'to make sensible the machinery of probability theory') jointly entail the VoIP can be viewed in McCarthy, Mikkola, and Thomas' (Theorem 2.3.1).[5]

One's theory of prudential value can make a big difference at this stage. I also omit those details here, but the implications of various positions (e.g., Strong Non-Comparativism or Comparativism) are rigorously teased out by Teru Thomas, and the interested reader should seek those answers there if they so wish.[6] At any rate, I have already commited us to Conditionalism and explained how this is consistent with our assumption of Existence Noncomparativism; so, I will press on.[7]

[2] As noted in an earlier chapter, we will write $\mathbb{L}(i)$ for the prospect faced by the i^{th} individual in some lottery $\mathbb{L}$ involving n persons.

[3] (Thomas, 2016, 121)

[4] These are described at (Thomas, 2016, 120-128).

[5] (McCarthy et al., 2016)

[6] See (Thomas, 2016, 130ff).

[7] If we instead adopted Existence Comparativism, then it would follow, on the Competing Claims View (specifically, under (i)), that a nonexistent person has an interest at stake (so, a claim) for having

You will remember, Conditionalism ranks prospects by their value conditional on existence. Paired up with the VoIP, this corresponds to evaluating welfare distributions as if we were evaluting for the sake of an individual whose identity is uncertain, but who is sure to exist. The paradagmatic axiological framework is therefore Averagism.[8]

But this isn't your father's Averagism. As Teru points out, there are a few extra wrinkles in the 'chancy' cases. In 'non-chancy' cases—this is to say, where the outcome of every lottery is certain—what we can call 'Veiled Averagism' (hereafter abbreviated 'VA') performs just the same as classic Averagism.[9] So, it is vulnerable to many of the better-worn objections commonly levelled against Averagism in population ethics. (Below, I will go on to discuss the top four which are particularly brutal.)

If the outcome of a lottery is uncertain, then we find that the VoIP pushes us away from the two standard formulations of Averagism: (a) expected average utility; and (b) average expected utility. Here, I'll just reproduce Teru's arguments since he has already put it best. Consider first, against (a), the following two lotteries:

$\mathbb{L}_{100}$	s_1	s_2
Ann	Ω	-20
Bob_1	10	Ω
Bob_2	10	Ω
...	...	...
Bob_{99}	10	Ω

$\mathbb{M}_{100}$	s_1	s_2
Ann	Ω	10
Bob_1	-20	Ω
Bob_2	-20	Ω
...	...	...
Bob_{99}	-20	Ω

If we take the expected average utility, then $\mathbb{L}_{100}$ amounts to

$$V(\mathbb{L}_{100}) = \sigma(\tfrac{10\cdot 99}{99}) + (1-\sigma)(\tfrac{-20\cdot 1}{1}) = \sigma(10) + (1-\sigma)(-20)$$

Now we determine the value of $\mathbb{M}_{100}$:

$$V(\mathbb{M}_{100}) = \sigma(\tfrac{-20\cdot 99}{99}) + (1-\sigma)(\tfrac{10\cdot 1}{1}) = \sigma(-20) + (1-\sigma)(10)$$

If $\sigma = 0.5$, then

$$V(\mathbb{L}_{100}) = V(\mathbb{M}_{100})$$

the alternative adopted which brings him into existence. This seems to me too strange to be even remotely plausible. While the alternative in which he exists, all else being equal, could be good for him as well as better overall in terms of the population's value, it surely cannot be true that he has a claim. He's just not the right kind of creature to bear interests or make claims on others.

[8] cf. (Thomas, 2016, 148)

[9] If an outcome A has a lower (higher) average lifetime welfare level than another outcome B, then A is worse than (better than) B. If they have the same average lifetime welfare level, then A and B are equally good.

But this cannot be right. $\mathbb{L}_{100}$ is clearly better than $\mathbb{M}_{100}$ from behind a veil of ignorance. It's more probable that you are a Bob than not, and Bobs are better off (if s_1 is the actual state of affairs) under lottery $\mathbb{L}_{100}$.[10]

A similar problem arises for average expected utility. This time, while being Ann or Bob is equiprobable, the probability of a bad outcome obtaining under one of the lotteries is not. Consider:

$\mathbb{L}'_{100}$	s_1	s_2	...	s_{99}
Ann	Ω	10	...	10
Bob	-20	Ω	...	Ω

$\mathbb{M}'_{100}$	s_1	s_2	...	s_{99}
Ann	Ω	-20	...	-20
Bob	10	Ω	...	Ω

Average expected utility cannot account for the fact that the value of $\mathbb{L}'_{100}$ is greater since the outcome with Ann walking out of the rabbit hole with 10 is most likely to obtain.[11]

Instead, in chancy situations we find that VA ranks lotteries by *expected total utility divided by expected population size.*[12,13] VA ranks $\mathbb{L}_{100}$ above $\mathbb{M}_{100}$ and $\mathbb{L}'_{100}$ above $\mathbb{M}'_{100}$.

However, note that VA has a rather curious feature—specifically, the size of a population within an outcome (within an uncertain lottery) can in some sense swamp. To illustrate, imagine the following three distributions are possible under some $\mathbb{L}$:

$$\text{A} = \{2,\ 2,\ 5,\ \Omega,\ \Omega\},$$
$$\text{B} = \{5,\ 5,\ 2,\ \Omega,\ \Omega\},$$
$$\text{C} = \{2,\ 2,\ 2,\ \Omega,\ \Omega\}.$$

VA describes the value of this lottery as being $3.\bar{2}$. But if we manipulate C so that it were instead $\text{C}^* = \{2, 2, 2, 2, 2\}$, then VA determines the value of this lottery to be ≈ 2.63. If we take the limit of the size of the population, C, to infinity, then the value of this lottery converges on 2.[14]

This will stick in some of our craws. Indeed, it bears extreme similarity to a well-known (negative) implication of Averagism in the non-chancy case. It, of course, behooves me to explain how it is that the contributive (dis)value of additional persons in C* maps onto the Competing Claims View I plan on developing in this chapter.

[10] It also violates Principle 2. See (Thomas, 2016, 150).

[11] It also violates Principle 1. See (Thomas, 2016, 150).

[12] (Thomas, 2016, 150)

[13] Our theory, VA, thereby acts on the Self-Indicating Assumption (or what David Lewis and Adam Elga have called the Thirder position (with reference to the Sleeping Beauty Puzzle)). This is not uncontroversial in anthropics. See especially (Bostrom, 2002); cf. (Elga, 2000); (Lewis, 2001). In the appendix I run the toy model on the Self-Sampling Assumption (or Halfer position) instead—just to see what pops out. Basically, I just crunched the numbers such that the value of lotteries was determined by average expected utility. As it turns out, the results are even stronger than what we will find on my current toy model.

[14] See the appendix for a brief survey of what happens if we allow for infinities to enter into our toy model.

After all, we can imagine there not being an alternative lottery, and if so, then their interests are at simply not at stake. Therefore, it seems that VA is fundamentally misaligned with the Competing Claims View.

In fact, besides the problems that I have already mentioned, there is yet more reason to doubt the compatibility of VA *&* the Competing Claims View. VA reduces overall value to impartial but personal value. The Competing Claims View starts off doing the same, but goes a few steps further. On the Competing Claims View, relational goods (e.g., equality) contribute towards the overall value of lotteries as well—*but only insofar as they strengthen a person's existing claim; mere inequality, for example, does not ground a claim when he cannot be any better off.*[15]

Basically, there is a tension between the two views: *VA implies stuff that the Competing Claims View doesn't, and the Competing Claims View implies other stuff which the VA doesn't.* Put simply, they are mismatched. In response, I'll argue that VA is nevertheless appropriate, and is better than any other account I have considered, for tracking the Competing Claims View in the context of variable populations.

The crux of my argument is that the Competing Claims View overlooks an important consideration of fairness, 'worthwhileness', which should also be applied to measure the strength of competing claims. I am unaware of anyone else that has made a similar claim in the literature. Yet, this extension of the Competing Claims View seems to me both natural and intuitively robust.

There are roughly two procedures for measuring the strength of competing claims. Following Simon Beard, it could be that the badness of certain aspects of unfairness combine productively. For example, Beard's suggestion is that it is unfair to be worse off than one might have been, *but it is even more unfair to be worse off than others through no fault of his own, except where one either could not have been better off than one is and so on.*[16]

This is not the kind of procedure I plan on adopting in my arguments. Contra Beard, I suppose that the strength of competing claims depend on several distinct elements of an outcome, some of which directly concern fairness, whose values one merely adds together. I go this route because: (a) it fits the language of (Fleurbaey and Voorhoeve, 2012) a little better;[17] (b) it seems to me the harder route (so, less prone to the charge of ad hockery); and (c) allows me to present some novel arguments to the effect that worthwhileness massively outweighs the collective force of the other elements. I'll argue that:

> (a) its relative weight grows without bound as we scale the size of the affected population up, while this simply isn't true of most other considerations; or
>
> (b) it outweighs those other unbounded considerations on a one by one basis.

[15] See (Fleurbaey and Voorhoeve, 2012, 397).

[16] (Beard, forthcoming)

[17] Indeed, it is difficult to make sense of claims such as " the *more* his interests are promoted by ..." on Beard's preferred framework.

My core claim is that once *worthwhileness* has been factored into our overall evaluations, we find that overall value just about totally reduces back down to personal value. Thereby, we will have successfully salvaged VA as the best proxy for the Competing Claims View (in large population cases).[18]

But do we really need such a proxy? A critic might scoff; according to him, at best it's overkill, and at worst I have just wasted everyone's time.

My critic is wrong. It's not obvious to me how someone plans on evaluating the kind of large-scale decision that I wish to explore (in Part 2) with the Competing Claims View. To begin, the math required for pulling out the Competing Claims View is going to be very complex in chancy cases spanning all possible human histories, given there will be lots of moving parts with nebulous values that need to be nailed down. Furthermore, VA isn't subject to one's own ideas about how to define the function describing this or that consideration's effect on the strength of a claim, whereas the Competing Claims View is. *Simpler is simply better.*

*
* *

Here is a prospectus.

The mismatch between VA and the Competing Claims View has three independent sources. Firstly, the VoIP is itself inconsistent with some of the considerations that the Competing Claims View incorporates into its assessment of overall value. The second and third sources of our problem I have already summarized: *VA implies stuff that the Competing Claims View doesn't, and the Competing Claims View implies other stuff which the VA doesn't.* More specifically,

> A. The VA implies some kind of swamping phenomena for outcome goodness in our overall evalution of a lottery as the size of a population grows larger.
>
> B. The Competing Claims View *could be understood* as lending more weight to the claim of a person that is worse off through no fault or choice of his own.[19] More so, a claim is strengthened only if (and to the extent that) the person experiences a shortfall under some lottery.

Section 4.2 rehearses the Competing Claims View, and unpacks the four considerations that prevent overall value being evaluated in terms of personal value. (Note that it only takes any one of the four on its own to do so.) One can immediately recognize that the three principles from which we derived the VoIP are incompatible with three of those four considerations that contribute to the overall value of lotteries according to the Competing Claims View.

(a) Equal chance of benefit is denied by Principle 1.

[18] I do not defend the claim that VA can substitute for the Competing Claims View in each (hence every) case that involves competing claims. If my arguments go through, they implicate only large population cases. Therefore, the VA so-described is intended for those types of really big decisions.

[19] See (Temkin, 1993).

(b) Equality is denied by Principle 2.

(c) Priority (and, so, (iib)) is rejected by Principle 3.

This is the greatest source of my headaches. I'll argue directly against the value of 'equal chance of benefit'. However, I postpone slaying the beast, as it were, until section 4.5.

Section 4.3 introduces the four most common objections to Averagism (in non-chancy cases). As such, VA implies some stuff that, at first blush, is inconsistent with the evaluations of lotteries under the Competing Claims View. Only some of these objections survive once we factor in the details of my toy model, as section 4.4 goes on to demonstrate.

Section 4.5 takes us all the way home. I begin by defining 'overpopulation', and argue that we have considerations of fairness that could be reasonably couched in claims made by persons. Afterwards, I argue that the swamping which occurs under VA is no blemish or indication of its inability to evaluate lotteries on the Competing Claim View's behalf. This is because 'worthwhileness' operates in a similar fashion so far as it weakens (or swamps) the other considerations down to nothing if we take the limit of the size of the population to infinity.

Section 4.6 closes the chapter by introducing Part 2.

4.2 Competing Claims View. II

Bear in mind, the Competing Claims View states that:

> *Competing Claims View*: we decide between alternatives by considering the comparative strength of the claims of different individuals, where
>
> (i) a claim can be made on an individual's behalf if and only if his interests are at stake; and
>
> (ii) his claim to have a given alternative chosen is stronger:
>
> > (iia) the more his interests are promoted by that alternative; and
> >
> > (iib) the worse off he is relative to others with whom his interests conflict.

Below, I list the four considerations which strengthen an individual's claim against some option being taken.

Right off the bat, though, I must highlight once more that unless someone's interests are at stake—meaning that, he would be better off under some alternative—none of the following considerations kick in. We can illustrate this with one of the four considerations which Fleurbaey and Voorhoeve do not explicitly incorporate into the Competing Claims View, but which they clearly think has some role to play.

Example. On this view, inequality is not intrinsically bad. Mere inequality, as such, does not ground a claim when a person could not be made better off by removing the

inequality. "Nonetheless, between people with competing claims, inequality matters because it lends force to the claims of those who are worse off."[20]

There is another consideration which functions just the same as equality on the Competing Claims View. This consideration, though, some of us might not want to say concerns matters of fairness even though it alters the overall value of lotteries. Reading (iib) the consideration of 'priority' is taken to be a relational good. Even if we are both very well off, benefits to me contribute more towards outcome goodness so far as I am worse off than yourself.[21] It is very different in substance from the more standard claim that the worse off someone is in absolute terms—so, non-comparatively—the more it matters that benefits go to him.[22]

The remaining two considerations, like equality, seem to sprout from a concern for fairness.

First, (iia) tells us that a person's claim is stronger the greater his shortfall under some lottery. This, you will remember, is going to be devilishly difficult to try and capture in non-simple toy examples involving large (variable) populations. Moreover, we can strengthen this consideration to also reflect a plausible, related aspect of fairness: luck. Specifically, let's say that a person's claim is even stronger if he is worse off than he might have been through no fault or choice of his own. Second, Fleurbaey and Voorhoeve endorse equal chance of benefit as strengthening a person's claim insofar as it contributes to the fairness of the situation.[23] In their own words:

> The intuition that giving people equal chances of being advantaged can make such a distinctive contribution to fairness appears to be widely shared. The best explanation of this judgement seems to us to be that a given outcome inequality among people with equally strong claims to a benefit is less unfair when each person has a chance to end up better off than when the worse off have no such chance, because in receiving this chance, each person receives something of expected value. Chances of receiving benefits, or of avoiding harms, are of expected value to people, and the more equally this value is distributed at a relevant point in time, the fairer the distribution. Of course, the *well-being value* of a chance evaporates once it is clear that this chance is unrealized. However, a chance's contribution to fairness does *not* evaporate. It remains true, for example, of someone who received an equal chance at an indivisible benefit, but for whom the benefit did not materialize, that he had a fair chance of receiving it.[24,25]

[20](Fleurbaey and Voorhoeve, 2012, 397)

[21]One of the more difficult puzzles that proponents of the Priority View have on their hands is making precise the notion of 'same size benefit': *under what conditions does one benefit accruing one person count as being the 'same size' as a different benefit accruing to a different person?* For a survey of the problem refer to (Greaves, forthcoming-*a*); for an overview of related issues see (Parfit, 2012) (and all articles published in the same special issue of Utilitas).

[22](Parfit, 1991, 19, 22); (Parfit, 1997, 213); (Brown, 2005, 201)

[23](Fleurbaey and Voorhoeve, 2012, 395-397)

[24](Fleurbaey and Voorhoeve, 2012, 396) (their emphasis); cf. (Wasserman, 1996)

[25]See also (Otsuka, 2017).

All four of these considerations contribute to the overall value of some lottery in ways that aren't reducible to sheer axiology. Indeed, as one of my toy examples brings to light, giving someone an equal chance of benefit may be worse for him in expectation. And this makes the VA a bad candidate, as such, for standing in for the Competing Claims View.

*
* *

Teru, in developing the VoIP, put forward a series of stochastic dominance arguments against some of these considerations. In fact, the first in this series he calls 'Against Fairness'! Because a single example will suffice, I will reproduce this particular argument below.

Some will recognize this toy example from the introduction of the dissertation. This time around, though, we will finish what we started. Consider the following two lotteries:[26]

$\mathbb{L}$		$\mathbb{L}_1$	s_1	s_2
Ira	x	Ira	x	y
Eli	y	Eli	y	x

Table 4.1: Diamond's Case

Under $\mathbb{L}$, Ira is certain to get x and Eli is certain to get y. By contrast, under $\mathbb{L}_1$, there is a half chance of Ira getting x while Eli gets y, and a half chance of Ira getting y and Eli getting x.

Now, it is true that behind the veil it does not matter to the interlocutor who gets x, and who gets y under either lottery. $\mathbb{L}$ and $\mathbb{L}_1$ are equally good so far as his expected welfare goes—in short, his chance of getting x is just as high on both lotteries. This might stick in someone's craws, as it in fact did for Diamond.[27] He maintains that $\mathbb{L}_1$ is sometimes fairer than $\mathbb{L}$ for the reason that the distribution is random, and not predetermined. This is sometimes referred to as *ex-ante equality*.

In his own thesis, Teru suggests that we might wish to reject this type of fairness for being non-axiological. In other words, it cannot be the case that what makes ex-ante equality good is that it is good for persons. To see this, let's follow Teru Thomas' suggstion of modifying $\mathbb{L}_1$ such that there is a small cost for picking this lottery: $\mathbb{L}_1^-$.[28]

$\mathbb{L}_1^-$	s_1	s_2
Ira	$x-$	$y-$
Eli	$y-$	$x-$

Table 4.2: Diamond's Case Revised

[26](Thomas, 2016, 122)
[27]See (Diamond, 1967).
[28](Thomas, 2016, 123-124)

Crucially, if the cost is small enough, then $\mathbb{L}_1^-$ could be even better than (or at least as good as) $\mathbb{L}_1$, the lottery we threw out. But consider now that the outcome of $\mathbb{L}_1^-$ on s_1 must be worse than $\mathbb{L}$, given that everyone is worse off. More so, the outcome on s_2 can only be as good as the outcome on s_1 insofar as they are mere permutations of each other.[29] Therefore, "Diamond's judgment suggests that $\mathbb{L}_1^-$ is no worse than $\mathbb{L}$ despite the fact that it is certain to have a worse outcome. (More broadly: despite the fact that $\mathbb{L}$ 'stochastically dominates' $\mathbb{L}_1^-$.) This is counterintuitive."[30]

This kind of response is clearly unavailable to me. After all, I have not set out to rip out the internals of the Competing Claims View and keep it in name alone. Put differently, I *do* accept that the value of a lottery can depend on (e.g.) proportional relations that do not reduce to how good an outcome is in personal terms.

Another tempting response is this.

My claim is conditional. VA can succesfully stand in for decisions involving lotteries which bear far more similarity with $\mathbb{L}_1$ than they do with $\mathbb{L}$. So, for example, we can safely ignore the effect of equal chance of benefit in Part 2 of my thesis. This is because, as I have said, there are too many possible histories of the world—on my model, as in real life, it only takes one small perturbation to a single variable to alter the course of history! This, to be sure, isn't to suggest that there is *some* consistent pattern to history. The earliest forebears are always, for example, cave-men on my toy model. But we don't want to take stock of a lottery that alters conditions during this period. It's, after all, a fact that cave-men had no shot at being much better off or colonizing space. To build such a lottery into the thesis would amount to losing grip on reality.

However, I don't think I need to concede anywhere near as much ground. I intend to show that my arguments hold across a range of other dilemmas in population ethics which might be of interest for advocates of the Competing Claims View.

My reply is instead this.

While stepping around Teru's argument, given that we are not only looking at how good an outcome is for persons, we could nevertheless learn a valuable, general lesson from his stochastic dominance argument. *Even if these non-axiological considerations matter, they can only contribute so much towards overall value.* In other words, an equal chance of benefit adds value, sure, but less than you might have thought.[31] After all, there is only so much of a cost we would be prepared to ask Ira and Eli to accept before $\mathbb{L}_1^-$ is worse than $\mathbb{L}$. We might go even further. Imagine if the cost was entirely shouldered by Eli (the person that's worse off in $\mathbb{L}$)? This distribution of the costs seems a little unfair even if it provides an equal chance of benefit. My tummy feeling is that this is less fair then if both Ira *&* Eli ate the cost. At any rate, it is not

[29] This move depends on the truth of John Broome's *Principle of Impartiality Between People* (Broome, 2004, 135). (Note that this principle goes by different names in the literature. For instance, Teru dubbs it *Anonymity* (Thomas, 2016, 117).) It is, at any rate, a cherished axiom that most of us are not prepared to give up. See, for example, (Frick, 2017, 356-358) for the argument that Broome should be prepared to accept 'greediness' on pain of rejecting 'impartiality'.

[30] (Thomas, 2016, 124)

[31] This, by the way, is already endorsed by Fleurbaey and Voorhoeve, as well as others (e.g., Simon Beard). See (Beard, forthcoming) and (Fleurbaey and Voorhoeve, 2012, footnote 31).

obviously unreasonable to assume that equal chance of benefit is worth less in these circumstances. Now suppose that another person, Leroy, who is sure to receive z, a piddling amount of welfare, under either $\mathbb{L}$ or $\mathbb{L}_1$ must pay the cost for choosing $\mathbb{L}_1^-$. It's hard to imagine that equal chance of benefit for Eli has much if any force left.[32]

This, of course, does not mean that the VA and Competing Claims View are suddenly compatible. Not yet.

All you need to accept at this stage is that equal chance of benefit, as well as the other considerations when properly-implemented, improves overall value by some finite amount. Notice too, the lower the effect (on strengthening a person's claim) that these considerations have, the more powerful my subsequent argument about worthwhileness and swamping.

Moving along, the next section introduces four particularly brutal objections against Averagism.

4.3 Four Objections to Averagism

Averagism has come under heavy fire within the field of population ethics in the past. The four strongest objections to Averagism are widely considered to be fatal.[33] These are:

(a) *Averagism will sometimes forbid happy mere additions*;

(b) *other times it will require (happy or miserable) mere additions;*

(c) *Averagism ignores how large the affected population size is when picking between (at least some) outcomes; and*

(d) *it will ignore the extreme suffering of a few as the size of the population grows.*

This is no storm in a teacup. Mere additions are persons that if brought into existence would in no way affect the welfare of the remaining population. So, their very presence

[32] Admittedly, our intuitions concerning priority to the worse off might be muddying the water here. In other words, perhaps equal chance of benefit does not lose any of its force here; rather, there is a competing consideration which is being set against it. But even if that were so, it still tells us something important. In a population comprising persons with competing claims, the balance could tip this or that way depending on the strength of every relevant individual's claim. We will find that this outcome is better only by some margin. For example, in a population of size 200, we can imagine 10 persons having a claim in support of ϕ while another 10 have a claim against ϕ. The scales will tip here by only some small margin. By contrast, worthwhileness is assessed by looking at the long-term trajectory of these 200 persons. And if it undermines these benefactors' claims, then it will undermine them even further the larger this population grows in size *even if only 10 persons have a claim against* ϕ. Obviously, I have yet to argue this point. But it's good for us to bear in mind in advance that the ratio of what's at stake between a claim in support of ϕ and what's at stake according to worthwhileness by ϕ-ing will grow (with worthwhileness eventually swamping the other consideration) as the size of the population grows.

[33] See (Parfit, 1984, 406, 420-22); (McMahan, 1981, 96ff); and (Parfit, forthcoming); (cf. Pressman, 2015). In personal communication Toby Ord has said of Averagism that "this is a view we can safely ignore". It's my understanding that John Broome also holds this position.

in the world cannot generate a competing claim by others. To solidify the complaint, we can also say of these persons that they could not be made better off—they are *uniquely realizable*. If that were so, then they themselves would have no grounds for a claim.

The Competing Claims View most definitely doesn't generate either (a) or (b).[34] And when it comes to (c) and (d) we can foresee that the Competing Claims View will be at least misaligned with Averagism on specific cases.

Let's begin. Below, I'll describe how each objection puts pressure in a different way on the idea of Averagism properly mapping onto considerations of fairness.

4.3.1 Obj. 1: *Egyptology*

According to Averagism, it is worse if the average lifetime welfare level of the (timeless) population is lower than it could have been. When comparing two outcomes this might not ruffle too many feathers. However, as Derek Parfit has argued, this generates a strange, counter-intuitive implication when it comes to happy mere additions. I quote his toy example in full:

> On the Average Principle, the best history might be the one in which only Eve and Adam ever live. It would be worse if, instead of Eve and Adam, a billion billion other people lived, all with a quality of life that would be almost as high. Though this claim is hard to believe, it is not absurd. The second history is in one way worse. It is bad that no one's life is quite as good as Eve and Adam's would have been.
>
> The Average Principle has other implications which *are* absurd. Suppose that Eve and Adam lived these wonderful lives. On the Average Principle it would be worse if, *not instead but in addition*, the billion billion other people lived. This would be worse because it would lower the average quality of life. *This* way of lowering the average, by Mere Addition, cannot be plausibly claimed to be had.[35]

As Parfit goes on to unpack, similar problems arise when we start factoring in the lifetime welfare of a mere addition's forebears or far-off progeny. Consider first the case of *Egyptology*.[36] If we are attempting to maximize the average happiness of the (timeless) population, then the goodness (or badness) of bringing someone into existence will depend on facts about all previous lives, including those of the ancient Egyptians. If most

[34] Suppose that under lottery $\mathbb{L}_1$ either (a) only a and b ever live at 20 lifetime welfare levels or (b) 100 other people live with lifetime welfare levels of 18. Under lottery $\mathbb{L}_2$, it is (a) or (b*). If (b*) obtains, then all hundred plus a and b will live. Suppose that the outcomes are equiprobable under either lottery. Now compare what the Competing Claims View and Averagism have to say. According to the Competing Claims View, both lotteries are equally good. By contrast, Averagism describes $\mathbb{L}_1$ as having 60 expected value, and $\mathbb{L}_2$ as having ≈ 19 value; therefore, $\mathbb{L}_1$ is better than $\mathbb{L}_2$. In short, it would be better to take a chance on one or the other group being alive, but not both.

[35] (Parfit, 1984, 420)

[36] (Parfit, 1984, 420); citing Jeff McMahan's paper, (McMahan, 1981, 96ff)

Egyptians had a very high lifetime welfare, then this counts against the goodness of bringing a moderately happy persons into existence now. But, as Parfit notes, "research in Egyptology cannot be relevant to our decision whether to have children".[37]

The case can be reversed.[38] Instead of looking backwards in history, we instead look to the lifetime welfare levels of our distant progeny. Depending on how well off these future people are, bringing a mere addition into existence now can be good or bad with resepct to the population's average lifetime welfare. If we imagine that in the far future the average lifetime welfare is extremely high, then the addition of a moderately happy person now would lower the average lifetime welfare of the (timeless) population. Again, I quote Parfit in full:

> It is then more likely to be bad if I have a child, even if my child's life would be well worth living, and his existence would be bad for no one. It is more likely that my child's existence would lower the average quality of all future lives. This cannot be relevant. Whether I should have a child cannot depend on what the quality of life will be in the distant future.[39,40]

In all three toy examples we see that Averagism prohibits us from causing happy mere additions to exist. But their lives would be worth living, and no one else would be affected by their existence.

This is not so with the Competing Claims View in a wide range of circumstances we could imagine. Suppose that we are deciding between two policies. On the first, humankind goes extinct on Earth because it didn't want to go exploring the dark vastness of space. On the other, humankind spreads out across many galaxies *but exploring the uncharted wild west of space is hard, and their quality of life is lower than our own.* In this decision, the mere additions have no claims to make; they would not be better off (or worse off) under the alternative—indeed, it would be incomparable so far as they were concerned. All else being equal, the Competing Claims View would not prohibit humankind from spreading out across many galaxies (but nor would it require it!). So, Averagism and the Competing Claims View disagree about happy mere additions in at least some circumstances.

4.3.2 Obj. 2: *Hell 3*

There is another sense in which Averagism holds that whether bringing a person into existence depends on irrelevant facts about other people's lives. Parfit asks us to imagine:

> *Hell 3.* Most of us have lives that are much worse than nothing. The exceptions are the sadistic tyrants who make us suffer. The rest of us would

[37](Parfit, 1984, 420)

[38]Indeed, Michael Pressman refers to this as *Reverse Egyptology* (Pressman, 2015, 399).

[39](Parfit, 1984, 421)

[40]If we imagine instead that in the far future the average lifetime welfare is extremely low, because some catastrophe has permanently ruined our world, then Averagism will say bringing the moderately happy person into existence makes things go better.

> kill ourselves if we could; but this is made impossible. The tyrants claim truly that, if we have children, they will make these children suffer slightly less.[41]

Averagism states that the world goes better if we bring these miserable children into existence. It doesn't matter that their lives would be terrible; this is irrelevant, given that their existence would raise the average lifetime welfare of the timeless population.

The Competing Claims View, while not prohibiting bringing these miserable lil' ones into existence, does not take a step towards the dark side and require it of us.

Consider instead if these tyrants offered to make some much smaller number of the same children much better off (and the remaining children remain merely possible). Depending on how many of us there are with lives much worse than nothing, the average could be highest by refusing the new offer for the former. But doing so leaves some number of children *worse off* (i.e., those that would exist in either outcome), and so in a position to complain that they have been treated unfairly.

There is a pattern here that some of you will have recognized. According to the Competing Claims View, we do not owe possible persons a happy life just because it would be happy. What we do factor in on this view is *if* we chose to bring this person into existence under certain conditions, *whether he would have been better off in some other state of affairs.* If so, then this shortfall must be set against whatever other values suggest this option is best.

Nor does it seem like we can justify the creation of miserable people on the grounds that others before them (or after them) were even worse off *but would not benefit in any way from the creation of these new miserable persons.*

4.3.3 Obj. 3: *The Two Hells*

A third objection often levelled at Averagism is that this axiological framework ignores (to some extent) how large the affected population size is when picking between (at least some) outcomes. The problem has been illustrated by Parfit:

> *Two Hells.* In *Hell One*, the last generation consists of ten innocent people, who each suffer great agony for fifty years. The lives of these people are much worse than nothing. They would all kill themselves if they could. In *Hell Two*, the last generation consists not of ten but of ten million innocent people, who each suffer agony just as great for fifty years minus a day.[42,43]

On Averagism, Hell One would be worse than Hell Two, since the lives of these ten persons would be slightly worse than the lives of the ten million in Hell Two. While

[41] (Parfit, 1984, 422)

[42] (Parfit, 1984, 406)

[43] Parfit often claims that a miserable life would be much worse than nothingness. This comports with our intuitions. However, I do not have recourse to such a claim in my arguments. After all, I claim that existence is noncomparable with non-existence. So, we must ignore this evaluative claim on the part of Parfit when going forward.

it is true in one respect that Hell Two is better, it's also surely true that in another respect Hell One is better. After all, "[in Hell Two] the total sum of suffering is nearly a million times greater. And it can be claimed that this vast increase in the sum of suffering morally outweighs the very small reduction in the sum of suffering within each life."[44]

It's not obvious to me that the Competing Claims View does not also go against the grain of intuitions here. So long as no one could be made better off, there are no grounds for them to claim that the alternative is better. And yet if the first ten persons exist in both outcomes, then they do have a legitimate (and unopposed) claim for Hell One.

Still, for completeness, we are at least aware of this objection.

4.3.4 Obj. 4: *Indifference to Torture*

The final objection I'll cover in this section breaks the camel's back, I think. If we were to take the limit of the size of the population as it approaches infinity, the addition of a miserable person whose life is not worth living—such that, he would kill himself if only he could—contributes zero to our evaluation of an outcome (as the averages converge).

This problem is worth laying out slowly so as to pick out where it counterbalances the severity of the earlier objections. To begin, notice that Averagism is concerned with how happy people are, not how much happiness there is in an outcome. It's not better for the world that persons exist because their lives contain happiness. Rather, "happiness is good because it is good for people".[45] The size of a happy population is largely a matter of taste on these grounds.[46] However, Averagism is merely *semi-neutral* about causing happy persons to exist.[47] After all, if we could bring into existence a person far happier than either ourselves or our forebears, then this would raise the average and, therefore, make for a better outcome.[48] Therefore, Averagism is not purely neutral with respect to the contributive value (in terms of outcome goodness) of adding happy persons to our population. But the contributive value of an individual diminishes as the size of the population grows. If we were to take the limit of the size of the population as it approaches infinity, then the addition of yet another happier-yet person would contribute zero as the averages converged.[49] As such, the previous objections are softened as we consider larger and larger (timeless) populations—e.g., the creation of a (single) moderately happy child will be swamped by the large size of the population, contributing nearly nothing to our evaluation of its goodness (or badness); if the (timeless) population is infinitely big, then this child's contribution is negligible.

[44](Parfit, 1984, 406); (cf. Pressman, 2015, 417-420)

[45](Parfit, 1984, 394)

[46](Narveson, 1967, 68)

[47]The *neutrality intuition* says that we are morally neutral with respect to mere additions (whether they be happy or miserable). See (Broome, 2004, 143ff).

[48]See especially (Shulman, 2014).

[49]See §5.3 of (Greaves and Ord, forthcoming-*c*).

But some might be, as I'm sure my dear readers in fact are, appalled by what this implies about the badness of creating lives not worth living. After all, Averagism implies that as the size of the population grows, the more we should tend towards indifference regarding the harrowing plight of the one child kept in perpetual filth, misery, and darkness in order to sustain their utopia.[50] Indeed, Averagism not only ranks an outcome in which a large enough number of people benefit from the torture of a single person as potentially better, but it will rank an outcome in which a single person is tortured for absolutely no gain (or loss) to other members of the (timeless) population as growingly irrelevant in terms of evaluating a growingly large population.

Again, it seems like the Competing Claims View runs into a similar problem here.[51] But there is a twist. Suppose that instead of keeping the child in perpetual filth, misery, and darkness in order to sustain their utopia, that they could have their 'Shangri-La' and release the child from his terrible fate too. They would have no claim against doing so since they would be equally well-off in both outcomes. So, the child's claim to be released would trump.

Averagism, on the other hand, provides, if at all, a growingly weaker reason to release the miserable lil' one from enslavement as the total population size grows larger. This is because Averagism allows large enough numbers to swamp the badness of some personal tragedy; the larger the size of the population, the more Averagism tends towards indifference regarding the child's plight.

*
* *

To summarize, there is reason to doubt the compatibility of Averagism and the Competing Claims View in the following four ways:

> (1) *The Competing Claims View doesn't forbid creating happy mere additions, all else being equal.*
>
> (2) *Nor does it require creating persons even when their lives would be so horrible they would rather kill themselves, all else being equal.*
>
> (3) *Averagism is (to some extent) insensitive to the number of persons affected by a bad outcome.*
>
> (4) *Either creating a miserable person or placing an individual into hellish conditions tends towards irrelevance on Averagism as the population size grows, but the same isn't true of the Competing Claims View (when all else is not equal).*

[50] I am, of course, alluding to *The Ones Who Walk Away From Omelas* (LeGuin, 2015).

[51] Although these implications may be very upsetting, it's old hat that population ethics is haunted by the impossibility of satisfying all of our intuitions. We are bound to run up against gnarly implications no matter which view we adopt. So, I do not take either this or the previous objection to be a decisive blow against the Competing Claims View.

4.4 Missed Their Mark

Some of these objections miss their mark, I'll argue. Once we take stock of the structure of our world, we just don't find that Averagism says some of these things. So, those inconsistencies (at least) between VA and the Competing Claims View go away. I'll then argue that those inconsistencies that do remain aren't actually problematic because the Competing Claims View's all-things-considered ranking meshes with that of VA.

**Of course, some of my critics will maintain that an axiological framework oguht to be able to get all the right answers—even for those colourful, highly artificial toy examples. I'm less sure about that.

Subsection 4.4.1 describes the sample space of outcomes which I think best fit the real world (to the best of our understanding). Section 4.4.2 then re-analyzes the above list.

4.4.1 The Structure of Our World

Some of this will come off as me being borderline condescending. Much to the contrary, I am merely being thorough.

4.4.1.1 Humble Origins

Every population originates on this planet at the same time in Earth's history. Furthermore, the starting conditions for intelligent life are rather gloomy. There is hunger, pestilence, and far more that mars the existence of the earliest persons. Our cave-dwelling forebears, we shall say, start off with lives just barely worth living in every possible outcome.

4.4.1.2 Grim Fate of Life

What we know of physical eschatology[52] tells us that life must come to an end. The indefinite survival of humanity, in other words, is physically impossible.[53] DEATH will come for us someday.

4.4.1.3 Doom & Gloom

The grim conclusion that life must come to an end is buttressed by the abundance of existential risks that plague all possible worlds, one of which we presently call home. Life could wither away rather than blossom. These existential risks present as 'filters'[54] or dangerous steps in the evolution from cave-dwelling to galaxy-colonizing civilization;

[52]This is the study of the future evolution of astrophysical objects (Ćirković, 2003b).

[53]See the appendix, *The $\mathcal{OP}$ is No Place for Infinite Ethics*, for a brief survey of what happens when the interlocutor takes infinite populations seriously. The same observations apply to VA.

[54](Hanson, 1998)

"[threatening] the *premature extinction* of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development".[55,56]

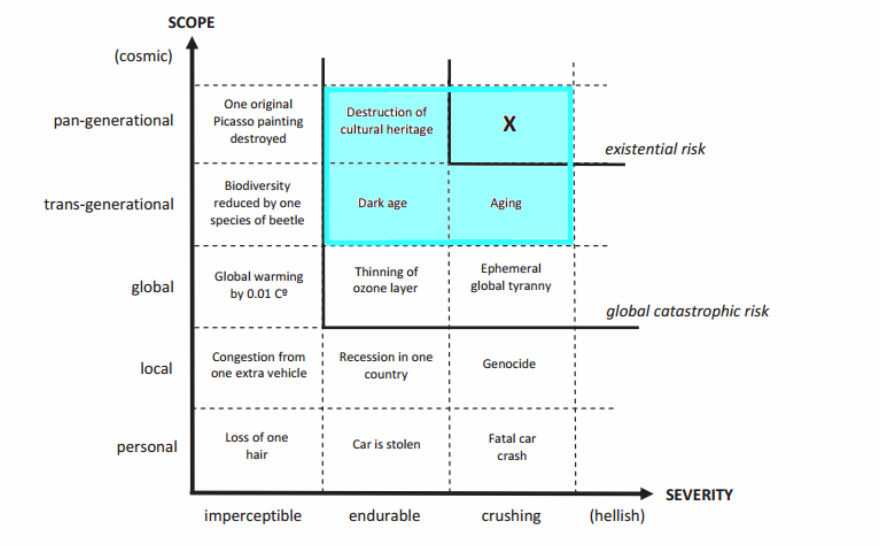


Figure 4.1: Qualitative Risk Categories - For our purposes, the possiblity of a biblical-sized threat that would be both cosmic and hellish is ignored behind the veil of ignorance. (Original source: (Bostrom, 2012a, 17).)

Existential catastrophes threaten to crush everyone's wellbeing post-event. There's no coming back from them. But things could also go very badly for a shorter period of time. Looking to the chart (in figure 4.1), I will assume that all worlds are plagued by the kinds of threats captured in the blue box. Threats which aren't at least trans-generational and endurable can be ignored. This is because I am solely interested in what we owe future persons. Harms that a sub-population might impose on itself or that are imperceptible are a topic for another dissertation. However, this being said, there is the possibility, of course, that ephemeral or small-severity threats may either (a) aggregate (e.g., carbon pollution) or (b) have indirect effects which snowball over time. Harms of this sort can culminate into a minimally trans-generational, endurable catastrophe.

I cannot here tackle perturbations to population trajectories from these kinds of harms—they are, in short, to complex to forecast or model accurately.[57] However, I

[55](Bostrom, 2012a, 15) (my emphasis)

[56]A fine-grained analysis of the different categories of existential catastrophe (and how they compare to, for example, global catastrophic risk) is available in the same paper by Bostrom.

[57]The subject of indirect effects is *especially* difficult. In Shelly Kagan's words, "lacking a crystal ball, how could you possibly tell what *all* the effects of your act will be? So how can we tell which act will lead to the best results overall—counting *all* the results?" (Kagan, 1998, 64); cf.(Lenman, 2000). Take the example of ridding the world of tuberculosis (TB). It's possible that its absence could be worse for the poorest regions of the world because of some unexpected ramifications of having done so—such as overpopulation or the next Hitler doesn't perish as a baby from TB. Contrarily, if TB is left to plague

do not think we lose anything by leaving them out of our analysis. After all, if the VA concludes that catastrophic risk ought to be minimized, then this will imply that we ought to take steps to also avoid these kinds of smaller harms, given they can give way to catastrophe.[58] The obverse is also true. If VA doesn't mind activities which raise catastrophic risk, then what bother is it if some smaller harms take place along the way? After all, the expected (negative) value of the former is likely much higher than the expected (negative) value of the latter in his deliberations behind the veil of ignorance.

There are three key points for us to bear in mind throughout this dissertation when it comes to catastrophes. Firstly, I make the assumption that no matter how life ends, it will be quick and painless. More so, I treat it as if members of this sub-population had lived a full life. This is done purely to simplify the toy model.[59] Secondly, humankind confronts not just the risk of extinction, but also the possibility of a broken world. Life could be monstrous, chock full of cruelties for centuries (or longer). He might be left out in the cold, slowly starving as the fabric of civility unweaves, breaking down into anarchy, fraught with suffering in an all-out-war of all against all. For instance, it could be man's fighting over oil in the darkest of the wolf hours which culminates in a nuclear winter that some tiny pocket of persons survive by hiding away for centuries in underground bunkers. Thirdly, I mean to draw a sharp distinction between the risk of (a) avoidable catastrophes and (b) freak catastrophes. Mankind has the power to reduce or increase their population's ability to manage or prepare for the former kind of threat. *The latter are, terrifyingly, out of their hands.* A broken world might be 'fated'. A single gamma-ray burst might do the dirty deed. Alternatively, a bubble of true vacuum could form, expanding throughout our cosmological horizon, disrupting our configuration of fields in Minkowski spacetime—a single four-dimensional entity in which all events occur—and go on indefinitely at close to the speed of light. Empty space, everywhere this bubble expanded, would suffer catastrophic gravitational collapse.

4.4.1.4 Our Long-Term Potential & Fragile Endowment

Despite starting off in rather grim conditions, things could continually get better if our forebears, us, and subsequent persons take the right steps.

The idea for how I might capture this in the toy model came to me after encountering Sir Partha Dasgupta's recent work. Dasupta asks us to imagine our planet as

the world, it could mutate into a highly-contagious super-bacterial infection that cannot be detected during the incubation period, culminating in a global pandemic. *Ex ante* it's difficult to tell if the combination of direct and indirect effects will be good or bad in the long-term for our population. And, as Hilary Greaves argues, the danger of (the new problem of) *cluelessness* is so much more pressing, real when we have good reasons for suspecting particular correlations between acts and indirect effects, but there are too many such connections pointing in different directions and no canonical weighing-up operation to aid us. See (Greaves, 2016). See also my reply to Hilary Greaves at (Kaczmarek, 2017, footnote 25).

[58] Indeed, this is precisely what I argue in Appendix A.

[59] In large-population cases it will not make any substantive difference anyhow, and this is precisely the kind of decision that my toy model concerns itself with in Part 2.

harbouring a finite amount of goods which is replenished in regular cycles by up to $\mathcal{K}$, "like manna from heaven".[60,61] For example, if left untouched, the rainforests will continue supplying a steady state of plants, animal life, water, and so on. If the earliest man devours more than $\mathcal{K}$, then his progeny will have less than him and $\mathcal{K}$ will be irreversibly lowered by the difference. So, even if successive sub-populations refrain from eating up more than $\mathcal{K}$, they will never be able to restore $\mathcal{K}$ to its initial number. However, if he eats up less than $\mathcal{K}$, then the world will not go on to produce an unbounded number of goods in the long run. There is a limit to how much the world can stably maintain (on its own). Therefore, saving $\frac{1}{2} \cdot \mathcal{K}$ per cycle will not lead to an extreme abundance for his progeny in the long run. And if he follows a plan of moderate subsistence, eating up exactly $\mathcal{K}$, then the amount of goods in his world will remain steady over time.[62] So, in a nutshell, Dasgupta has presented us with a *killing the goose that lays the golden eggs* scenario that is both realistic and probable.[63]

My own model goes a different direction, but is similar in spirit to Dasgupta's model. For starters, my description of outcome goodness is split between (*i*) how one's forebears spend energy (from their finite total energy pool) and (*ii*) the development path his forebears will take. By comparison, Dasgupta's argument runs only on a single variable: how a sub-population might affect $\mathcal{K}$. My reason for splitting up the interlocutor's concern is so that we can handle the problem of population policy indepedent of how much we ought to sacrifice to mitigate catastrophic risk. How much energy our forebears deplete affects our population's *long-term potential*; meanwhile, our forebears' development path affects the *endowment* we stand to inherit.

Although they are distinct issues, both concern what we owe each other. The matter of population policy concerns how we ought to distribute some finite amount of energy.[64] *While we do not owe possible people a happy life just because it would be happy (on the Competing Claims View), the very potential for life itself does fall within the domain of fairness. After all, it seems to me that a future generation can coherently claim that we took more than our fair share of the energy pool. Meanwhile, the matter of our long-term development path concerns how we ought to distribute the harms (or burdens) associated with a world plagued by catastrophic risk. After all, persons living in Tim Mulgan's broken world can rightly complain that they have been unfairly left with an onerous share of the harms corresponding to a catastrophe which we could have*

[60](Dasgupta, forthcoming-*a*, 6)

[61]For simplicity, let's say that these goods are both countable and commensurable.

[62]Dasgupta's set-up also includes a discursive function which factors in our interest in seeing our progeny do well (which factors in their interest in seeing *their* progeny do well, and so on) which bears similarity to Rawls' own thinking on the matter. See (Dasgupta and Dasgupta, forthcoming-*b*); (cf. Rawls, 1971, 284ff).

[63]I say probable here because a world may not be plagued by radical scarcity if $\mathcal{K}$ is depleted at some earlier epoch. After all, the population may become technologically mature and no longer depend on the planet itself to replensih their food supply (e.g.).

[64]To be sure, I do *not* deny that there are renewable sources of energy in our world. There are. But this would complicate the model terribly. Plus, in the big scheme of things, sources of energy like this are in relatively short-supply. I assume, at any rate, that the finite energy pool my toy model draws on already factors in renewable energies.

prevented by sacrificing some elements of our affluent lifestyle.

I plan to use the term 'long-term potential' in a non-teleological sense. It is proportionate to the total energy pool available to a given population. This provides a final theoretical limit on how much energy a population could utilize. Depending on the rate at which living creatures consume energy in a given population, the timeless population can (a) advance only so far; (b) grow only so large; or (c) last only so long. Indeed, the three are intertwined. To illustrate, the larger the population is at an earlier stage of history, the less energy there will be for bringing about future persons—so, unless they trim down their numbers down the road, the population will starve from energy-depletion sooner.

Let's say that the average living creature's total consumption starts off uniform across all possible worlds, and the consumption rate of a single living creature is fixed.[65] This is the strongest limit I'm going to bake into my own toy model. Bear in mind, according to Dasgupta's model, the population can only consume as much or grow only as large as 'manna falls from heaven' (before they permanently ruin $\mathcal{K}$). According to my own toy model, a single living creature has a relatively small impact and cannot exhaust the energy pool on their own. Still, I'll leave it open to the earliest cave-dwelling persons in history to propagate like no rabbit ever before has, and *collectively* devour the entire energy pool. Furthermore, if they were to do so, then their lifetime welfare would not drop accordingly. I'm going to allow, in other words, that their lifetime welfare is orthogonal to how large this sub-population happens to be.

These two features of my toy model are, of course, preposterous. But though this is definitely absurd, I think we ought to be all the more confident of the verdict reached on my model if there's no limit baked in that punishes the Go-Fast policy for choosing to eat up more than $\mathcal{K}$. On my toy model, the extinction of mankind sooner rather than later is not intrinsically worse. Nor can persons impoverish themselves simply by overpopulating the world at any temporal segment.

So far I have discussed how a population's energy pool will directly influence (a), (b), and (c). But there is the further issue of how different development paths will increase or decrease subsequent generations' endowments.

If my own toy model were to have adopted Dasgupta's suggestion of $\mathcal{K}$, then VA would recommend adopting a sustainable rate of growth (or savings) rather than a sustainable trajectory. In this way Dasgupta's argument strongly resembles the arguments from Rawls. Both want to establish a rate of growth—where Dasgupta has population size in mind, and Rawls has savings in mind—that is to some extent fixed. In Dasgupta's case, this rate of growth is optimal if a sub-population could only be made better off by disturbing $\mathcal{K}$. In Rawls' case, the rate required is capped by what it takes to make institutions a permanent feature of our civilization. Even taken together, these two rates of growth fail to account for the breadth of problems plaguing the real

[65] Although there might be fluctuations here and there throughout history (e.g., cave-men will have a poorer metabolic rate (in terms of energy consumption) than agents residing in a Matrioshka Brain), and amongst different agents within the same sub-population (e.g., the wealthy may consume more than the poor because they have better access to the energy pool), we will make the simplifying assumption that their consumption rate is ergodic unless otherwise stated.

world. It will not be enough to merely limit population size or leave behind enough for future people. Taking the rights steps so that humanity doesn't get stuck in the mud of some gloomy cave also requires *steering* our development path through uncharted, perilous territory.

As I have earlier said, the earliest persons in human history start off in rather grim conditions. A novel feature of my own toy model is that persons can initiate *eucatastrophes*[66] which improve their own lifetime welfare level. I will hereafter abbreviate 'eucatastrophe' as $\mathbb{E}$. These moments in human history are the dopplegängers of catastrophic events. An $\mathbb{E}$ is a change of conditions in the world for the better. $\mathbb{E}$s have enduring effects on the development trajectory of the population. But not merely because they have indirect effects that snowball over time. As before, we will ignore these kinds of nebulous effects. Rather, what I have in mind are long-run events that generate inexhaustible benefits which are available at all times to a population after being caused to exist—e.g., finding a cure for smallpox.[67] Furthermore, I assume that, *ceteris paribus*, persons from all subsequent epochs benefit equally from long-term eucatastrophes.[68]

The development path of his forebears dictates how many $\mathbb{E}$s a person inherits. If they go fast, then he will inherit more. And the more $\mathbb{E}$s that he has access to, the better off he is on my toy model. But he also stands to lose. The promotion of $\mathbb{E}$s is dangerous. If they are haphazardly created by going quickly, and skimping on safety precautions, then cumulative catastrophic risk goes up. Indeed, we frequently see this take place in real life (e.g., the race to build artificial intelligence).[69] This being said, though, going fast on my toy model also involves what I'll hereafter refer to as *quick-and-dirty* $\mathbb{E}$s—e.g., there might be more to gain by advancing their stock of culinary

[66] As pointed out in (Cotton-Barratt and Ord, 2015, 3), "[the term] was coined by Tolkien to refer to the sudden and unexpected turn for the better frequently found at the end of fairy tales."

[67] The concept is derivative of Toby Ord's and Owen Cotton-Barratt's "Existential Eucatastrophe": an event which causes there to be much more expected value after the event than before (Cotton-Barratt and Ord, 2015, 3). The key difference between 'eucatastrophe' and 'existential eucatastrophe' is the magnitude of the benefits; on the former, small victories count too.

[68] I will assume equality among persons belonging to the same generation. Furthermore, I cannot here tackle how aggregation within a lifetime ought to be done (see esp. Broome, 1991). This is an important topic, but it would take me too far afield to try and cover it in sufficient detail. Instead, I will assume that the development trajectory tells us how many $\mathbb{E}$s a sub-population has access to, and this in turn tells us everything relevant about the randomly sampled person's welfare in that generation. So, if a person p_1 is in population A, then p_2 is in A if and only if p_2 has the same welfare as p_1.

[69] As Nick Bostrom observes,

> [We] humans are like small children playing with a bomb. ... For a child with an undetonated bomb in its hands, a sensible thing to do would be to put it down gently, quickly back out of the room, and contact the nearest adult. Yet what we have here is not one child but many, each with access to an independent trigger mechanism. The chances that we will *all* find the sense to put down the dangerous stuff seems almost negligible. Some little idiot is bound to press the ignite button just to see what happens. Nor can we attain safety by running away, for the blast of an intelligence explosion would bring down the entire firmament. Nor is there a grown-up in sight (Bostrom, 2014, 259).

and literary goods than by either slowing down the ageing process or curing malaria.[70] After all, the latter are going to very difficult by comparison. Abiding to this kind of schedule means ignoring the risks imposed on people in the far future. The goal, to repeat, is to improve the world around them as quickly as possible so that they themselves will benefit. By contrast, if our forebears develop slowly, then they will reduce the threat of (cumulative) catastrophic risk that we will face. The population will be able to survive for longer. But this will mean that the endowment eacb of us stands to inherit would be smaller. After all, safety-engineering the world against those monsters lurking along our development path from cave-men to galactic-colonizers will mean delaying the creation of other $\mathbb{E}$s.

Having said this, there are some similarities between Dasgupta's model and my own. For starters, a sub-population's endowment gets passed on to subsequent sub-populations (like his $\mathcal{K}$). And, like Dasgupta's model, a sub-population may permanently jeopardize the endowment of future sub-populations if they make some poor decisions along the way.[71] $\mathbb{E}$'s *aren't* a permanent feature of the world once caused to exist. For at least some of them, their presence in our world depends on certain pre-conditions staying intact. Moreover, there may be $\mathbb{E}$s which open the door to new threats we would never have otherwise come across—a pandora's box scenario. To illustrate, messing around with cocktails of ancient diseases (e.g., bubonic plague) in some government-run organic chemistry lab could, in effect, reintroduce a deadlier, mutated strain of small pox into the world. Worse, this fragment of civilization may no longer have the ability to manage the infectious disease if the solution has long been lost in the dustry corridors of decrepit libraries from a time long gone.[72] To sum up, just because we may have cut the head off of one monster (e.g., age of pestilence) doesn't amount to having slain one of the four horsemen of the apocalypse. Some of the monsters we will fight along the way are hydras.

Because it will facilitate a smoother discussion down the road, let's distinguish falling short of their long-term potential from shortcomings in the endowment passed

[70] Of course, ridding the world of some deadly pathogen (e.g.) requires a few biologists to roll up their sleeves, and put in long hours at the lab, as well as perhaps denying themselves some personal happiness (e.g., love for the game of baseball). Persons might prefer to consume their own energy pool for activities which don't involve creating more $\mathbb{E}$s—e.g., a freerider who watches baseball instead of curing fatal diseases. My toy model doesn't attempt to account for the freerider problem even if we might have learned something worthwhile about what we owe each other by digging into the matter. It would only bedevil my project with unnecessary technical complications which we can safely ignore. After all, by my own design, this sub-population can only benefit itself by creating more $\mathbb{E}$s. To illustrate, they are surely better off if they watch baseball without being in the grips of a malaria-induced fever. It doesn't matter that they must bear the brunt of bringing these $\mathbb{E}$s into existence. Nor will it matter to this sub-population that the benefits of having done so must be shared equally with all subsequent sub-populations down the road. Again, we are only considering goodness for persons, and part and parcel of being well off is creating $\mathbb{E}$s themselves.

[71] Contrary to the actual historical record, where our cave-dwelling ancestors really couldn't do *all that much damage* to the world, we shall also say that the earliest cave-dwellers are capable of leaving behind a barren, terrifying wasteland if they were to choose to eat up their world.

[72] In this case we might say that the necessary pre-conditions for keeping smallpox under wraps deterioriated.

on from forebears to progeny. Let's call these, respectively, 'waste' and 'rot'. Because all possible worlds are plagued by catastrophic risk, and sloppy mistakes are bound to happen as well (we are after all just human), there is no population trajectory which has zero waste or rot. The goal, as such, is to determine which balance of the two is least harmful in terms of every person's expected lifetime welfare.

To illustrate, it might be that a person is better off if his forebears exhaust the energy pool and take a dangerous but quick development path. Bear in mind, the VA corresponds to evaluating welfare distributions as if we were evaluting for the sake of an individual whose identity is uncertain, but who is sure to exist. The randomly sampled person should thereby reason as follows about our case: *if* his forebears were to exhaust the energy pool and take a dangerous but quick development path, *then* he would (very probably) find himself at the beginning of history with a high lifetime welfare level.

4.4.1.5 Humanity's Resilience: *Brave Pioneers of the Wild West*

It seems reasonable to assume that mankind's tenure among the stars could outstrip that of the Earth's. Although their initial energy pool is limited to what the Earth produces, we will suppose that any population might find a way to survive the hailstorm of astrophysical processes ahead, spread out across several galaxies, and go so far as to find sources of energy in, for example, black holes well after the era of stellar evolution has passed.[73] In short, the only thing holding a population back from persisting under very extreme conditions is the current state of their scientific prowess. If they act wisely, then Earth-originating intelligent life may survive for millions, billions, or even trillions of years, and spread out very far in space.

Now, insofar as this is supposed, we must recognize that it's possible for a population to be spread out across a supercosmological horizon. That is to say, along the way, we might spread out in many directions as we colonize the supercluster of galaxies, such that, given the growing acceleration of the universe, subsets of the population are hived off along multiple cosmological horizons—leaving parts of the population causally disconnected from each other. But even so, because there is a causal connection between forebears and their cosmologically-isolated progeny, we will nevertheless refer to them as a single population.

4.4.1.6 The Folly of Anti-Natalism

Finally, my toy model will consider outcomes in which we go intentionally extinct only if this is done by either exhausting the energy pool or growing old and dying. I'm going to ignore, in other words, outcomes in which a sub-population kills itself (e.g. releases a deadly pathogen).

[73] A good survey of these processes and eras in physcial eschatology is offered in (Adams, 2008). There one can furthermore find reference to some of the challenges we will have to overcome (in applied cosmology) if we plan to survive them. Along the way, sources of possible energy include condensed molecular matter, degenerate stars, intergalactic gas, and dark matter halos.

This ought not ruffle too many feathers.[74] After all, in real life, not every member of the sub-population will want to go extinct. They'll fight it as hard as they can, raging against the dying of the light.[75] Imagine that the fallout of a thermonuclear war leaves a few surviving tribes peppered around the world. Even if the going gets tough, such that they are starving, cold, and miserable, it is hard to imagine that every tribesman would converge on ending it all then and there. It only takes a few poor souls to lose their nerve.

Moreover, it is very difficult to kill off humanity in the modal sense. According to Duncan Pritchard's account of risk, there is a difference in how risky bad outcomes are even if they are equally probable. Compare the following:

> *Case 1:* An evil scientist has rigged up a large bomb, which he has hidden in a populated area. If the bomb explodes, many people will die. There is no way of discovering the bomb before the time it is set to detonate. The bomb will only detonate, however, if a certain set of numbers comes up on the next national lottery draw. The odds of these numbers appearing is fourteen million to one. It is not possible to interfere with this lottery draw.
>
> *Case 2:* An evil scientist has rigged up a large bomb, which he has hidden in a populated area. If the bomb explodes, many people will die.There is no way of discovering the bomb before the time it is set to detonate. The bomb will only detonate, however, if a series of three highly unlikely events obtains. First, the weakest horse in the field at the Grand National, Lucky Loser, must win the race by at least ten furlongs. Second, the worst team remaining in the FA Cup draw, Accrington Stanley, must beat the best team remaining, Manchester United, by at least ten goals. And third, the queen of England must spontaneously choose to speak a complete sentence of Polish during her next public speech. The odds of this chain of events occurring are fourteen million to one. It is not possible to interfere with the outcomes of any of the events in this chain.[76]

As he goes on to argue, the first case is much riskier than the second case. While the first bad outcome requires very little change in the real world in order to obtain, the second describes a *far-off possible world* in which a significant number of changes must take place in the real world. And the further away this possible world is in terms of its *closeness*, the less risky it is, according to Pritchard. A few coloured balls falling

[74] I provide further reason to cast this sort of anti-natalist plan aside in the appendix, *Life After Extinction*. Roughly, there is some chance that life will re-emerge post-catastrophe. This raises two crucial considerations. Primarily, the total extinction of the (timeless) population is even more difficult than we initially imagined. More so, the anti-natalist threatens imposing grave harms on this sub-population. E.g., they might start off in worse conditions than those of our own cave-dwelling forebears, and their prospects for climbing out of these miserable conditions might be ruined.

[75] An allusion to Dylan Thomas' famous poem, *Do Not Go Gentle Into That Good Night.*

[76] (Pritchard, 2015, 441)

in a certain configuration places the first bad outcome within the space of the closest possible worlds. By contrast, the three events required to unfold for the second bomb to detonate are "incredibly far-fetched".[77] In his own words, "none of them is an event that could very easily occur. For all three to obtain would require an incredible run of events. That's not to say that there is no risk of the bomb going off, since all three of these events are genuine possibilities—as we might say, stranger things have happened. But the point is that the possibility that the bomb goes off in case 2 is not something that could very easily occur in the way that it is in case 1, even despite the sameness of the probabilities involved."[78]

It seems to me that a plan to kill off humanity is more like Pritchard's second case than it is the first case. There are a number of conditions that would have to be altered in the real world in order for such a plan to be adopted by every member of the sub-population. And several more far-fetched conditions are required before this gruesome plan for humanity's doomsday could be pulled off. In the closest possible worlds, there would be resistance to this anti-natalist plan. More so, different historical segments will place different limitations on the success of such a gruesome plan—e.g., size of population and access to doomsday devices.[79] Finally, in the real world, the bomb does go off if we pull the trigger on a doomsday device of this sort and fail (to go extinct). We will break the world if things don't pan out as intended.

If Pritchard is right,[80] then I am no longer sure how we ought to account for the risk of a broken world obtaining instead of the extinction of humankind after pulling the trigger on this gruesome anti-natalist plan. It's not, in other words, that it would bedevil my toy model with technical complications to try to factor in the probability of the two outcomes conditional on an anti-natalist plan. Rather, there is no way to determine from behind the veil of ignorance how *far-off* the possible world in which we do go extinct (or break the world) is from the real world. By contrast, going fast or going slow is a small change to the real world, and both are approximately the same distance from the real world.

*
* *

To summarize the structure of the world that my toy model presupposes:

[77] (Pritchard, 2015, 442)

[78] (Pritchard, 2015, 442)

[79] Having said this, I concede that there are circumstances in which a well-hatched anti-natalist plan (even if unilaterally executed) is more likely to succeed. For example, in some possible worlds the population will be tiny (at at least some temporal segment). It's a lot easier to kill off humanity when there's only a few hundred persons roaming a relatively small region. And it'll get even simpler once the anti-natalists no longer have to rely on sharpened sticks or cannons to execute their gruesome plan. Indeed, there will be a sweet spot in the evolution of mankind: *after we have matured into a development stage where doomsday devices are commonly available, but before we have spread out across a supercosmological horizon.* But while this weakens my cause for assuming that intentionally going extinct is really difficult at all times in history, it does not change the fact that it will be very hard for some, if not most of history. I'm grateful to Patrick 'Paddy' Miller for flagging this point.

[80] See especially (Yang, forthcoming) for staunch criticism of Pritchard's model.

> Human life starts off in rather grim conditions, and it cannot evade DEATH's gleaming scythe forever. In between the path is poorly-lit and perilous, and in that relentless dark there will be very many monsters patiently waiting for us to miss a step. If we make poor decisions, life could be horrible. More so, the only way the population will go extinct is if it exhausts the energy pool, grows old and dies, or fails to survive an (unintentional) existential catastrophe.

But don't panic just yet. There's reason to hold out (existential) hope;[81] indeed, the chickens may never come home to roost. If we make wise decisions, then humanity may survive for billions of years or more, spreading out in multiple directions across space. And once they do, the threat of an existential catastrophe wiping out humanity will go down.[82] Life could be wonderful.

Of course, there's no telling from our armchairs what gread goods we might uncover down the road. Not only can we can now grasp the substance of what Parfit calls 'the best things in life', but we simply cannot be sure that there are such goods to be found.[83] I'm going to make two rather controversial assumptions at this point in the master argument. I assume there are such things that are 'the best things in life', and that "[some] of our successors might live lives and create worlds that, though failing to justify past suffering, would give us all, including some of those who have suffered, reasons to be glad that the Universe exists".[84,85] I furthermore suppose that we have only scratched the surface of the best things possible in human history.

We are blind to some other things too. We simply have no idea how rocky the transition from cave-dwellers to masters of the universe will prove to be—that is, existential risk may be objectively very high (much higher than we know) and tricky to deal with or not. We just can't be sure. So, my toy model considers the full range of possibilities.

My toy model runs together the randomly sampled person's aim to organize the (timeless) pouplation in a way that benefits him most (in expectation) with his aim to preserve his inheritance. There are many different combinations available in pursuit of this dual goal. And none of these combinations can escape producing at least some waste or rot.

The two main policies we are considering are: (a) go dangerously fast; and (b) go safe-n-slow. Of these two development policies, the former requires the most explana-

[81] Toby Ord and Owen Cotton-Barratt define this as the chance of an existential eucatastrophe (Cotton-Barratt and Ord, 2015, 4)

[82] But bear in mind that the threat of a freak catastrophe of mammoth proportion (e.g., a bubble of true vacuum forming) prohibits existential risk from dropping all the way down to zero.

[83] "These are the best kinds of creative activity and aesthetic experience, the best relationships between different people, and the other things which do most to make life worth living" (Parfit, 1986, 161).

[84] (Parfit, 2017)

[85] As Michael Plant and I like to say, it passes the 'Parfit Test'. Our rough rule-of-thumb is that if Derek Parfit considered it plausible (or, better yet, wholeheartedly endorsed the view (as he actually did in personal communication!)), then it in fact is plausible whatever our own intuitions might be.

tion. I postpone this discussion until the next chapter. The latter is relatively simple. Safe-n-slow requires the population to minimize catastrophic risk. This role can be performed by directly targeting these threats to humanity's survival and prospects. But this also requires being more careful while promoting $\mathbb{E}$s—e.g., refraining from developing synthetic biology until we fully understand its consequences. I assume that there is a point in history (though I imagine this to be far down the road) when the former role is less pressing. As we grow into maturity as a civilization, persons will be wiser and more capable of forecasting dangerous applications of some $\mathbb{E}$, as well as preparing for and averting its associated threats. Furthermore, I assume that a safe development path is at least twice as slow as going dangerously fast.[86]

Developing $\mathbb{E}$s safe-n-slow pays off most in a world that is plagued by a tiny bit of (cumulative) catastrophic risk and doesn't get rocked by a freak catastrophe along the way. But the longer the (timeless) population lasts in the world the longer they are exposed to the threat of a freak catastrophe either wiping out humanity or dramatically ruining their quality of life[87]—but note that a freak (or man-made) catastrophe which they survive will only temporarily cripple their civilization as part-and-parcel of going safe-n-slow is safety-engineering the world in anticipation of such catastrophes.[88] If catastrophic risk is very high, then going fast might be the winning ticket—even if the cumulative risk of an existential catastrophe goes up. After all, they could rapidly improve their own welfare and then pull the plug—by producing lots of babies (and so exhausting the energy pool) before catastrophe strikes—right before everything came crashing down around them.[89]

[86]This is arbitrary, undoubtedly. However, I struggled to find a function which does a better job and did not require some rather ad hoc maneuvers on my part. Again, I assume that humankind will growingly have the power to make their world good. If so, then their ability to safely develop new $\mathbb{E}$s may speed up over time. (*Note:* this is not the same thing as their default development rate (e.g., logarithmic). However we define *that* function, we can transform it to properly reflect a slower or faster development approach.) And there will be different rates of safety-engineering required of us at various temporal segments in history. Cave-men probably just have to avoid burning down their villages or being consumed by a volcano eruption; so, can largely get on with the business of inventing the wheel (e.g.). Meanwhile, we have to avoid releasing a deadly pandemic among other things; and this seems far more time-consuming. After factoring all of these unknowns into my decision, I made my best guesstimate using the method of Fermi Estimates by taking the geometric mean of an estimated bound. This resulted in an average development speeed of half the default for safe-n-slow. After sanity-checking, this does not strike me as wildly unreasonable.

[87]Bostrom (correctly) distinguishes a state risk from a step risk. Freak catastrophes and (predictable) astrophysical threats (e.g., gamma-ray bursts or supernovae explosions) fall into this category. Bostrom describes a state risk as "the total amount of risk to which a system is exposed is a direct function of how long the system remains in that state" (Bostrom, 2014, 234). (See the appendix where I outline Ćirković's notion of a hostility parameter—where the basic thrust of his claim is that we are bound to approach an "equilibrium state in which perturbations from past large-scale physical processes (like nucleosynthesis and gamma-ray bursts) [cease] to play a significant role describing the transition between small and large civilizations (Ćirković, 2012, 75).) A step risk, however, is different. It is the discrete risk associated with some transition (Bostrom, 2014, 234)—e.g., the transition from a mature civilization capable of manipulating atoms to a wise civilization that doesn't threaten dropping atom bombs on each other in a game of chicken.

[88]See especially (Beckstead, 2015) and references therein.

[89]Forecasting the timing of a catastrophe will, of course, be remarkably difficult. Certainly, it doesn't

The full sample space of outcomes we are left with is far richer than the following suggests, but every outcome falls roughly into one of these four categories: go fast and suffer a broken world; go fast and early extinction; go safe-n-slow and early extinction; go safe-n-slow and survive for eons (with the occasional freak catastrophe temporarily lowering their welfare). As I have said in *Defeating the Lil' Monster in All of Us*, there is no way to know which outcome (determined by an action and a state, $o[a, s] \in \mathbb{O}$) will result by adopting either policy. There are countless ways things might get coloured in.

4.4.2 Averagism Revisited

Two features that I have built into my toy model require us to reformulate the concerns of Parfit and McMahan regarding the application of Averagism as an axiological framework. First, mere additions come at *some* cost on my toy model. Every living creature consumes some (even if tiny) portion of the total energy pool. Although it might well still be the case that the creation of a life does not in any way affect the welfare of the existing population, his presence in the world does affect subsequent sub-population's access to the total energy pool. Second, by creating more $\mathbb{E}$s, the gloomy conditions of our cave-dwelling ancestors can be left in the dust. After some stretch of time, life could be wonderful.

The four rough categories I presented above can be further broken down into nine more finely-grained outcome-types which reflect the different ways we might break our world.

(a) *The (timeless) population survives for a very long time and life continues getting better at every temporal step.*

(b) *At some early stage in history there is a catastrophe:*

- (b.1) *life goes extinct;*
- (b.2) *the world is irreversibly broken;*
- (b.3) *the world is broken but gradually recovers; or*
- (b.4) *the world is broken and gradually gets worse.*

(c) *At some later stage in history there is a catastrophe:*

- (c.1) *life goes extinct;*
- (c.2.) *the world is irreversibly broken;*
- (c.3) *the world is broken but gradually recovers; or*
- (c.4) *the world is broken and gradually gets worse.*

strike me as something we could do accurately in practice. I am nonetheless prepared to imagine, for the sake of argument, that practitioners of the dangerously fast development policy are very competent in this regard.

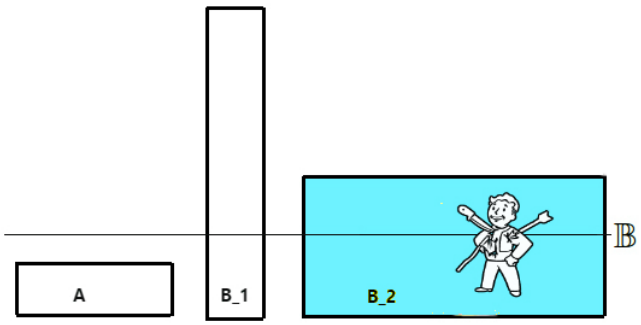


Figure 4.2: *Sometimes Miserable People Make Things Better.* The boxes represent sub-populations, and their height describes how well off the sub-population is, while the width tells you how many persons exist in that sub-population. Box B_2 is the population that we could bring into existence at a later, more menacing period in history.

Let's suppose that the population is uniformly distributed in every outcome. More specifically, because their total energy pool is finite, we assume that there is an even split of energy given some long-term potential. For example, if the population could survive 10^6 generations, and host 10^8 living creatures in total, then there are 100 living creatures per generation.

If the temporal distribution of the population is evenly split in this way, then Averagism will rank (a) as better than every other outcome. The matter of (b) and (c) is a bit more complicated. Let's stick to (b) for the moment. If we suppose that a broken world results in a lower lifetime welfare level than that of the earliest cave-dwellers, then (b.1) is better than (b.2.), and (b.2.) is in turn better than (b.4). Depending on the numbers—that is, how poorly off persons are at different temporal segments *and* how long convalescing takes before the population regains and surpasses their pre-apocalyptic lifetime welfare level—(b.3) is definitely better than both (b.2) and (b.4) but might be worse than (b.1). Again, this all hangs on how the numbers work out.

For the purpose of making this point clear, the case we are imagining can be coloured in by considering a population that is approaching the End of Days. There is still some energy left in their total energy pool—so, we could add more people. However, the upcoming 400 year period of history will be foreboding, dark, and menacing. The persons we bring into existence would be far worse off than we are, but their lives will still be worth living, and, indeed, they won't be as poorly off as their ancient relatives from the agen of gloomy caves and fierce beasts. Suppose their decision is between bringing these less well off persons about or growing old and dying.

A back-of-the-envelope calculation can reveal that adding more persons to a development trajectory which careens downward might nevertheless raise the average

lifetime welfare level in the timeless population. This isn't a new drum I'm banging. Carl Shulman has discussed this with reference to what he calls 'inaccessible populations'.[90] Looking to the figure, the inaccessible (sub-)population is A and B_1. We cannot go back in time and add more happy persons. Let $\mathbb{B}$ stand for the threshold of a life worth living.[91] Furthermore, let $\bar{X}$ stand for the average lifetime welfare of members belonging to sub-population X. If A has a low-enough lifetime welfare (defined by the size of their basket of $\mathbb{E}$s), then there is some value of $\bar{B_2}$ that raises the average lifetime welfare of the timeless population. Note that $\bar{B_2}$ is a function of not just how steep the drop is over time but the size of B_2 as well. If either too many of them are caused to exist or the drop in lifetime welfare is too harsh, then the overall numbers pan out a different way. Nevetheless, we can conclude that—though with every terrifying step the road gets a little darker, a bit more perilous—Averagism will rank (b.4) as better than (b.1) if:

$$\bar{B_2} > \left(\frac{\mathcal{N}(A) \cdot \bar{A} + \mathcal{N}(B_1) \cdot \bar{B_1}}{\mathcal{N}(A) + \mathcal{N}(B_1)} \right) \tag{4.1}$$

Of course, if growing old and dying is worse than adding B_2, then it must be true that exhausting the energy pool by enlarging the size of B_1 is best.

All of my above comments are the same for (c). If we compare (b.1) to (c.1) and (b.2) to (c.2), and so on, we find that it's better if the catastrophe occurs later in history according to Averagism. From the armchair we can also foresee that there will be some cases in which (b.3) is better than any of version of (c).

We can now ask ourselves if there is ever a time when an uneven temporal distribution of the population would raise average lifetime welfare. Well, if the size of the timeless population is held fixed, then every additional person that we create during an earlier development period will mean that one fewer person can be brought into existence at a later stage. Therefore, the average lifetime welfare will drop. Therefore, Averagism will rank this outcome as worse than the uniformly distributed (long) population.

This suggests that trimming down the size of the sub-population before the last generation is better. After all, the fewer miserable or moderately well-off persons there are, the less they will contribute towards the average. More so, this will allow the last generation, which is best off, to grow larger. Therefore, the average lifetime welfare of the (timeless) population will go up. Therefore, Averagism will rank the outcome in which it's iterations of Adam and Eve until right before the end as being better than a uniformly distributed (long) population. The same goes for populations that don't last as long. It is better if the less well-off members of the population are fewer in number, while the best-off members of the population are larger in number. Moreover, it applies in those cases where catastrophe is looming. When applied to variable populations, Averagism ranks the outcome in which the generation right before a catastrophe is

[90] (Shulman, 2014).

[91] I denote this threshold as $\mathbb{B}$ after Ramsey's *bliss level* for a (albeit perverse) reason that I'll unpack in a later chapter.

largest as better than the outcome in which it is equal in size to each other (past or future) generation.

There is a lower-bound to how small a generation can be, however, on my toy model. After all, underpopulation magnifies the threat of catastrophic risk. For example, a bad case of the flu could kill Adam, and thereby drag the whole of the future of humanity into the grave with him. Indeed, the grisly onslaught of (*inter alia*) pandemics and famines produces the need to keep our population's numbers up. Morever, the development path will be slower the fewer persons there are to conceive of and bring about $\mathbb{E}$s—and this could be astronomically wasteful.[92]

There are a few more caveats still. Notice that there's some lower-bound on how few persons can exist in every sub-population before the timeless population, in effect, fails to consume all of their available energy pool. Relatedly, it may be that some pockets of the total energy pool are trapped within some earlier cosmic epoch(s).[93] To illustrate, there may be a black hole burning itself out by the process of Hawking radiation. If this energy could only be utilized by a single generation, then there is *some number* of extra persons we could add which would raise the average lifetime welfare in the timeless population. Still, all else being equal, the size of each sub-population—apart from the final generation—ought to be minimized in keeping with the lower-bound, according to Averagism.

I'll now place the counterexamples in their proper pigeon-holes.

Every counterexample that refers to mere additions fails to apply on my toy model. Again, even if a life would be worth living, and no member's lifetime welfare would be affected by their existence[94], there is a finite upper-bound on how large the (timeless) population can be. There's no free lunch so to speak—every addition means that one fewer member can be subsequently brought into existence. In slogan form: *Egyptology does matter.*

Moreover, there is no outcome in the sample space which contains an infinitely big population—so, we can disregard (to some extent) the torture case. Finally, we can ignore Hell 3. On my toy model, the earliest cave-men start off with lives just barely worth living. This doesn't mean that Averagism won't rank an outcome in which we create miserable persons as better than going extinct. But it will only do so if there is an eventual and sufficiently strong upswing in terms of these miserable persons' own progeny's lifetime welfare level. If there are only more miserable persons on the far horizon, then Averagism will rank extinction as better than creating miserable persons.

This leaves all three versions of Egyptology and Two Hells for me to address. Let's start with Two Hells. Parfit objects to Averagism insofar as it would rank Hell One as worse than Hell Two, since the lives of these ten persons would be slightly worse than the lives of the ten million in Hell Two. This assessment inexplicably ignores the size and average lifetime welfare of the sub-population before it. Yet, two items

[92]This isn't meant to be a staunch normative criticism. Given how I have defined 'waste', this is indeed a very wasteful outcome.

[93]I'm grateful to Michael Plant for flagging this point in personal communication.

[94]This is a stretch, of course. If future generations have to impose strict population control, then someone is bound to be upset.

are immediately relevant to our assessment. When did this catastrophe take place in history? More so, how well off are their forebears?

If Two Hells were to take place early in history, then it falls under (b). The catastrophe responsible for their miserable conditions either affected some of their forebears as well or it did not. If it didn't, then Averagism will claim that it is better to bring about Hell One. If most of their forebears are as badly off as the persons in Hell One, then Averagism will claim that Hell Two is better than Hell One. If their forebears are better off or even more miserable than the persons in Hell One, then Averagism's ranking does not change: Hell Two is better than Hell One. In every case this is because the larger size of the sub-population in Hell Two will swamp the average lifetime welfare of their forebears.

If Two Hells were to take place later in history, then it falls under (c). Here, though, we find that in both cases—whether or not their forebears were poorly off—Averagism is growingly likely to rank Hell Two as worse than Hell One. This is because as the number of their forebears grows larger, the less an additional miserable person contributes towards the average lifetime welfare of the timeless population. By bringing about Hell Two we would threaten the miserable subset of the timeless population to either swamp back or swamp even more.

Let's move on. In all three vignettes of the Egyptology counterexample, the goodness (or badness) of bringing someone into existence will depend on facts about all previous and future lives, including those of the ancient Egyptians. If the ancient Egyptians had a much higher lifetime welfare, then this counts against the goodness of bringing a moderately happy person into existence now. Again we must ask ourselves how it came to be that present people are currently worse off than their forebears. There are three kinds of post-catastrophic histories available on my toy model. Either their world is irreversibly broken or lifetime welfare will either go up or down as more time passes. If the world is permanently spoiled such that every subsequent person brought into existence will be equally miserable, then Averagism will rank the outcome in which we bring them about as worse than going extinct after the Egyptians. This is also true if things are gradually getting worse in the world. If instead things would gradually get better, then prolonging the human race by bringing about persons that are worse off than the Egyptians might be better than going extinct—again, this depends on how the actual numbers pan out.

In Reverse Egyptology our progeny are much better off than we are. Here, we find that *if* no catastrophe (from which we could recover in sufficient time) were to rock our world, then adding more moderately happy persons now is worse according to Averagism. This would lower the average, as Parfit and McMahan note. But, bear in mind, the average also goes down even further insofar as every additional person brought into existence at an earlier stage in history comes at the cost of not being able to bring a future (happier) person into existence. There is an unavoidable tradeoff. Turn Reverese Egyptology on its head, and we find that *if* there were a catastrophe in the future (from which we could not recover in sufficient time), then adding more moderately happy persons now makes for a better outcome than the outcome in which

we failed to do so. Of course, Averagism will rank this outcome as better than the outcome in which we instead go extinct.

These are the rankings (concerning Two Hells and Egyptology) that I'll now argue, are consistent with what is governed by the Competing Claims View once we supplement it's assessment of overall value with the further consideration of 'worthwhileness'.

4.5 The Matter of Fairness Fleshed Out

Based on what I said in section 4.3, we concluded that there is reason to doubt the compatibility of Averagism and the Competing Claims View in the following four ways:

> (1) *The Competing Claims View doesn't forbid creating happy mere additions, all else being equal.*
>
> (2) *Nor does it require creating persons even when their lives would be so horrible they would rather kill themselves, all else being equal.*
>
> (3) *Averagism is (to some extent) insensitive to the number of persons affected by a bad outcome.*
>
> (4) *Either creating a miserable person or placing an individual into hellish conditions tends towards irrelevance on Averagism as the population size grows, but the same isn't true of the Competing Claims View (when all else is not equal).*

Subsection 4.5.1 will argue that, in fact, (1) is false. For starters, things are not equal on my toy model as section 4.4 described. Moreover, we *do* have claims against our forbears overpopulating the world. Section 4.5.1 concludes with me arguing that the Competing Claims View also implies (3). In the absence of any ill-will or recklesness by one's forebears, and all else being equal, it does not make a bad outcome any worse just because it would affect, not 5 or 500, but 5^{500} persons. This leaves us with both (2) and (4). In sections 4.5.2 *&* 4.5.3, I'll argue that these points, respectively, are mistaken about the Competing Claims View.

4.5.1 Overpopulation

I'm going to start by arguing that the matter of fairness does, under the right conditions, prevent us from adding more happy people. Specifically, I maintain that adding more happy people is prohibited by fairness if it threatens to overpopulate the generation. A generation becomes overpopulated if the addition of the last member placed any subsequent generation in the trilemma of choosing between (a) having to lower their own size in order to maintain an equal distribution going forward; (b) exhausting the remaining pool of energy by maximally propagating; or (c) leaving the next generation with an even harsher cut to make (in terms of (a)).

I'll make my case by appealing to the metaphor of colleges sorting out their allocation of rhubarb pie at dinner galas.[95] Suppose there is a massive rhubarb pie being handed down from dinner gala to dinner gala. Let the dinner gala represent a generation; the number of seats made available to dinner guests is the size of the generation; and the host of the dinner gala is the interlocutor acting on the behalf of these dinner guests.

> *The Distribution of Rhubarb Pie at Oxford.* Let's imagine that the provost cannot be fussed to deal with the small-fry business of rhubarb pies. So, he has organized a lottery system—the host of each college has written his college's name on a slip of paper, and entered it into an opaque box. The slips were then pulled out one-by-one, deciding the order in which dinner galas are to be hosted throughout hilary term. Suppose now that the entire rhubarb pie is delivered to a college for the dinner gala that evening. Let's say that an equal split between the colleges results in 50 dinner guests each. The host cannot change when they sit down to dinner. This boiled down to brute luck. But his fellow hosts and he can influence whether there's enough rhubarb pie to go around by deciding on how many seats to make available at their respective galas.
>
> Importantly, if earlier dinner galas consume all of the rhubarb pie, then no dinner guests will have their evening ruined. This is because no dinner guests are aware of the university's plan to supply a rhubarb pie. If they happen to get some, they'll be happy. But no host is going to spoil their evening by informing them that he intended to serve rhubarb pie until the greedy gentlemen of (e.g.) Magdalen College ate it all. Instead, the host will simply serve another dish which is incomparable in taste so far as the guests are concerned—indeed, the French dish of '*Nonèxistencę*' is a classic. However, if any of the pie remains for consumption, then the college must serve it—even if this means hosting a gala for one.
>
> The host currently in possession of the rhubarb pie can see how much of the pie is remaining, and is informed by the delivery-man of how many upcoming dinner galas remain in hilary term. Suppose that previous hosts invited more than 50 dinner guests on average. The current host must decide whether he will: (*i*) make 50 seats available for his dinner guests (if possible); (*ii*) make fewer seats available so that the difference is split between his college and subsequent ones (note: this needn't be an equal split); or (*iii*) invite as many dinner guests down as it would take to finish off the rhubarb pie.

My own reaction is that is was unfair of the previous host(s) to place the present host in this trilemma. This makes the outcome worse than one in which an equal split obtained.

[95] I've been told this metaphor isn't particularly helpful for some readers. While this may very well be true, I nonetheless find it easiest to spell out all the details of my own account/argument by appealing to such a metaphor.

Furthermore, it seems to me that if the present host chooses either (*i*) or (*ii*) under an uneven split,[96] then he too acts unfairly. But I don't think he is being unfair if he were to perform either (*ii*) under an even split or (*iii*) instead. After all, if the former, then every subsequent college can still have a gala. And if the latter, then, for starters, none of the dinner guests at subsequent galas are worse off; the food served in the place of rhubarb pie is incomparable in terms of its goodness. Plus, all 50 per college will get to 'eat nonèxistencę'. Therefore, I take it that the outcome corresponding to (*ii*) under an even split or (*iii*) is more fair than those corresponding to (*i*) or (*ii*) under an uneven split. And all four are less fair than the outcome in which either an equal split of 50 dinner guests per dinner gala obtained or a previous host simply served up all of the rhubarb pie (thereby, leaving no subsequent host in the trilemma).

Let's now add a wrinkle to the case. Suppose that rhubarb pie gets better with age, and the longer it ages, the better it tastes. The hosts of each college's dinner gala want their guests to be as happy as possible, and their guests would be happiest if they were having rhubarb pie at the very end of hilary term. But, to repeat, this is out of the host's hands. The timing of their gala has been decided by a lottery. At most the host of the gala can affect how many seats to make available for his dinner guests. Everything else about the case is the same. My sense of what fairness demands in this case does not get altered by the fact that dinner galas later in the term will benefit more from the rhubarb pie. They cannot, on grounds of fairness, demand that the hosts of earlier galas shorten their guest lists so as to save more of the rhubarb pie for later galas.

It is, of course, not unfair if they choose to do so. But they do not owe it to the guests of the later dinner galas in hilary term. Just because their share of the rhubarb pie won't have aged as long, so won't be as good, doesn't refute the fact that they have an equally strong claim to the rhubarb pie.[97]

Now, rhubarb pie isn't the same thing as a life worth living. But I think the case speaks for itself. Adding more happy people is prohibited by fairness if it threatens to overpopulate the generation. A generation becomes overpopulated if the addition of the last member placed any subsequent generation in the trilemma of choosing between (a) having to lower their own size in order to maintain an equal distribution going forward; (b) exhausting the remaining pool of energy by maximally propagating; or (c) leaving the next generation with an even harsher cut to make (in terms of (a)). This leaves open, you will recognize, devouring the remaining energy pool as fair even if none of your forebears have strayed from a fair distribution across time. This seems right. We do not owe possible persons a happy life just because it would be happy on grounds of fairness. Possible persons just aren't the right kind of entities to make claims or bear interests. Furthermore, devouring the rhubarb pie doesn't seem unfair to previous dinner galas. After all, they could have devoured the pie themselves. It's not in the

[96] E.g., imagine there are 10 colleges left to go. If he has 58 slices of rhubarb pie at hand, then he could serve 48 of his dinner guests, leaving either 10 for the next dinner gala or 1 slice per remaining dinner gala.

[97] Well, more precisely, in the language of the Competing Claims View it is true that everyone has an equally strong 'non-claim'.

hands of the present host to undo their choice to leave behind enough and well of the rhubarb pie for subsequent dinner guests.[98]

Turn the case on its head and the difference in direction (temporally) doesn't seem to make any difference. Suppose that rhubarb pie gets worse with age, and the longer it ages, the worse it tastes. The hosts of each college's dinner gala want their guests to be as happy as possible, and their guests would be happiest if they were having rhubarb pie at the very beginning of hilary term. But, to repeat, this is out of the host's hands. The timing of their gala has been decided by a lottery. At most the host of the gala can affect how many seats to make available for his dinner guests. Everything else about the case is the same. As before, my sense of what fairness requires does not get altered by the fact that dinner galas earlier in the term will benefit more from the rhubarb pie. They cannot, on grounds of fairness, leave the hosts of later galas with less rhubarb pie for their own guests.

Now we are in a position to consider (3). If the alternative is 'nonèxistencę', then the Competing Claims View will remain indifferent between outcomes where rotten rhubarb pie is served to different-size dinner galas. To see this, let's suppose that the rhubarb pie is in never-ending supply, but tends to go rotten at random. If the rhubarb pie goes rotten, it stays rotten. The last dinner gala is about to start, and guests are just sitting down to the table. The pie goes rotten. The outcome in which there are 50 seats reserved for the gala, and their rhubarb pie goes rotten, cannot be better than the outcome in which there are 5^{500} seats reserved for the gala, and their rhubarb pie goes rotten. The first 50 are just as badly off in both outcomes. No doubt, there are more dinner guests that are affected by the suddent bitter turn of the rhubarb pie on their plates. This might be bad for them. But it is not worse for them. And insofar as they have no claim to make against their forebears (after all, the rhubarb pie going rotten was an unpreventable freak event), the Competing Claims View should stay silent.

Summary. Overpopulation does generate a claim on the part of future persons. And since these additional persons' non-existence is incomparable, they have no grounds for the competing claim that they are better off under the overpopulation option. Therefore, (1) is false. Of course, a claim against overpopulation requires there actually being future persons. If our sub-population instead overpopulates and then goes extinct, then there would be no such claims made against our doing so. But this qualifcation does nothing in terms of salvaging (1) from the wreckage. Moreover, if all else is equal, the Competing Claims View is insensitive, like Averagism, to the number of persons affected by a bad outcome.

[98] Yet, there is one respect in which we might find this outcome to be unfair. It disrespects the wishes of the previous dinner gala hosts (for there to be more dinner galas). But this respect does not seem to transfer over in the case of population policy. We do not believe that we do something either wrong or unjust by disregarding our parent's wish for us to bear their grandchildren. This is my life, and I ought not be saddled by my parents' dreams of becoming grandparents if I would instead prefer to become an astronaut. (*I come back to flesh this toy example out shortly.*

4.5.2 Swallowing Our Rotten Pie

I now turn my attention to (2) *&* (4). My argument breaks down into three parts. To start, it will be shown that it is better, according to the Competing Claims View, if our forebears did not suffer the burdens of mitigating the risks plaguing our world if we just plan on going extinct soon anyhow.

We will consider two more vignettes of the rhubarb pie toy example. In both cases, it is assumed that the rhubarb pie gets better with age, and the longer it ages, the better it tastes.

In our first vignette, suppose there is an evil god, Moloch, that demands oblation be offered to Him in the form of a moderately big slice of rhubarb pie at every dinner gala. If a dinner gala refuses to make Him this offering, then he will slowly turn the remaining rhubarb rotten. Moloch is a bit unpredictable though. There's a small chance He'll do it even if oblation is paid in full. If the rhubarb pie goes rotten, then it stays rotten.[99]

Before I said that so long as the dinner gala's host doesn't presently find himself in the trilemma, then devouring the rhubarb pie is not a less fair outcome. He doesn't owe it to either subsequent or previous dinner guests to preserve rhubarb pie for future galas. I now want to claim that this verdict gets overturned if previous dinner guests had to offer up a spoonful (or more) of their rhubarb pie to Moloch in effort of salvaging the rhubarb pie for dinner galas later in the term.

This being said, our first vignette does not show this. At most we can glean that these past dinner guests had an obvious claim against paying oblation. It would have been a better outcome if previous dinner guests had not taken on the burden of placating Moloch. Past persons would have been better off, and everyone else would have been no worse.[100] My claim right now, then, is *merely* that the outcome in which they did pay oblation and the rhubarb is intentionally devoured sooner rather than later, is worse than the same outcome except they did not placate Moloch.

*
* *

Moving forward, does it make the world go worse for the present dinner gala to

(a) devour the pie, leaving nothing behind for subsequent galas; or

(b) send the pie along but either:

(b1) refuse to pay oblation to Moloch themselves; or

(b2) placate Moloch?

[99] In this case we are imagining that the rhubarb pie going slowly rotten represents the growing threat of cumulative catastrophic risk. His random act of rotting the pie represents a freak catastrophe.

[100] This seems to me importantly different from disregarding the wishes of previous dinner guests. To continue on my previous example, it does seem a little unfair to my parents to swear off having their grandchild if my parents have invested heavily into my child's college fund, taken lots of courses on handling or feeding infants, and so forth. But, again, it might be more unfair on me to bear the cost of having this child if it means impoverishing my ground projects (e.g., joining NASA).

At first blush, (a) cannot make the outcome any worse, given it is not worse for subsequent dinner guests.

By contrast, both (b1) and (b2) threaten to make things go worse. (b1) decidedly contributes to the badness of the outcome insofar as, on this lottery of theirs, subsequent dinner galas are worse off in expectation. *But*

> (i) subsequent dinner gala guests' shortfall may not be as great as the cost of now placating Moloch.

Still, even if that were so, this outcome is still worse than (a) insofar as either the present dinner guests or subsequent dinner guests are worse off than they could have been, and this is not true of (a).

Indeed, there is another way in which (b2) may contribute more towards the badness of the lottery than (a), given that

> (ii) Moloch might ruin the pie either way, and all they would be doing is making themselves worse off without benefiting any subsequent dinner galas.

Essentially, even if subsequent dinner gala's shortfall is greater than the cost of oblation on the present dinner guests, serving up rotten pie may be unavoidable.

I think this initial assessment of ours is wrong. Let's start with (i). I defend the axiological claim that (b2) could be at least as good as (a) and (a) is better than (b1) (so, it follows that (b2) could be better than (b1)). The next section speaks to (ii). There, I'll argue that it is false that (a) cannot make the world go worse.

The present dinner gala most certainly has a claim against paying oblation. But if they did pay their spoonful, this burden could be worthwhile. I will argue that the more worthwhile their burden, the weaker their claim against doing so.

My argument is grounded in a relatively simple idea.[101] The strength of one's claim against picking up the burden (placating Moloch) depends in large part on what possibilities this shuts down. Suppose that rhubarb pie didn't just get better with age, but it got better in ways that we cannot even now imagine. Dinner guests could someday be treated to the very best thing in life. Of course, it is regrettable that the present dinner guests will never experience such a wonderful culinary experience. It contributes badness (at least so far as it is the absence of goodness) to the outcome even if the present dinner guests could not have been better off in this respect.

And it does seem that offering up their spoonful or more of rhubarb pie to Moloch contributes even more badness to the outcome insofar as they are also worse off than they could have been (i.e., shortfall). This is, after all, what we are committed to on the Competing Claims View.

But I think the converse point applies in this context. It does not make things overall go much worse that they could have been better off by not sacrificing a spoonful

[101] Specifically, it is Derek Parfit's brief suggestion, here and there in the literature, of the 'best things in life'. For a very nice write-up of this relatively ambiguous position, see especially (Persson, 2017).

of their own rhubarb pie. For in order for them to be better off, they would have to risk taking the best things in life away from subsequent dinner guests.

The burden of being worse off than one could have been is worthwhile in this sense. It preserves the possibility of the very best things in life. More so, it can be maintained that displacing the burden onto subsequent dinner guests by not placating Moloch cannot be made worthwhile in the same way. The burdens subsequent dinner guests suffer are for nothing more than a tradeoff with their forebears.

To be clear, my claim isn't that the best things in life can justify previous suffering. Parfit did not find purchase in this idea, and neither do I. Outcome goodness is still initially determined by how good it is for persons. And the burdens of (*inter alia*) placating Moloch are going to count against outcome goodness in this sense.

Rather, the thought is that the additional considerations which strenghten a person's claim against some option get muted to the extent that they preserve the best things in life. This isn't to suggest that these considerations are morally unimportant. Contrarily, they are very important. But when they are set against the best things in life, they lose badly. Indeed, in Parfit's words, "what is best has more value—or does more to make the outcome better—than any amount of what is nearly as good".[102,103] Consider his toy example:

> *Century of Ecstasy.* 100 years that are filled to the brim with the best things in life after which humankind goes extinct.
>
> *Drab Eternity.* An infinite amount of time in which life is worth living but though "there is nothing bad in this life, the only good things would be muzak and potatoes".[104]

In terms of their objective value, it isn't hard to "share his view that of these two futures, the Century of Ecstasy would be more valuable, though the total quantity of value that it contains would be finite as opposed to the infinite quantity of the Drab Eternity".[105]

There's some danger here of me being misunderstood. I am not out to defend the idea that a Century of Ecstasy after some long period of human suffering would be far more valuable than the outcome in which persons were very well off in this long period and then went extinct. The 'best things in life' are not some kind of trump card, and the former might well not be more valuable than the latter.

Rather, I merely maintain that the extent to which persons have a claim against being worse off in the former is weakened by the fact that it means humanity losing out on the best things in life. This reduces their claim back down to nothing. Since, the persons in the Century of Ecstasy have no such grounds for a claim—this is to say, because they are uniquely realizable (and so, the alternative is incomparable)—we would then resolve the matter of how to rank these outcomes the old-fashioned way; by

[102](Parfit, 1986, 164); cf. (Persson, 2017, 106)
[103]For related discussion on aggregative procedures of this sort, see especially (Brown, forthcoming).
[104](Parfit, 1986, 160)
[105](Persson, 2017, 104)

turning to how good these outcomes are for persons. This is just what the VA tells us to do. And in those terms (of how well off each person is), the latter might be ranked above the former.

It follows that, if all else is equal, then (b2) is at least as good as (a)—and, therefore, better than (b1)—given the worthwhileness of their oblation to Moloch. Thereby, (4) is false. On the Competing Claims View, the shortfall of a miserable person does tend towards irrelevance so far as the additional considerations go.[106]

4.5.3 Best Things in Life, Swamping, & Freaks

This last observation allows me to finally challenge (2). To this end, I'll argue that under the right conditions (a) could be worse than (b2), as well as that (a) could be worse than the combination of (b2) and (ii).

This is really just a minor (and obvious) extension of the above argument. If our own claim against placating Moloch can be weakened, such that it does not contribute anything beyond how badly off we are to the overall badness of a lottery, then it is false that (a) is at least as good as (b2) in all circumstances. This is because just as our claim is weakened, so too is our forebears' claims against this alternative.

If we were to go extinct in our most recent rhubarb pie vignette, then while we would not introduce any further burdens into the world, we would have done either nothing at all (if there were no best things in life) or much less to offset the burdens shouldered by our forebears (if there were already some of the best things in life). So, (a) would still have some number of strong claims against it, and (b2) may eventually have either only very weak or outright impotent claims against it.

In a nutshell: We can make the world go better by prolonging humankind *but only if previous persons paid oblation.* So, we were wrong to say that (a) cannot make the outcome any worse in any circumstance.

The argument extends not just backwards, but forwards too. Consider now my second vignette.

Let's say that rhubarb behaves oddly. Sometimes it just goes rotten without warning. Knowing this in advance, every dinner gala is presented the opportunity to offer oblation to Gnon[107], and depending on the size of the offering—which must be paid in rhubarb pie (e.g., a spoonful from each dinner guest's plate)—Gnon will alter the chemical makeup of the rhubarb pie. While this cannot prevent rhubarb pie from going suddenly rotten, for it is a volatile thing (and Moloch doesn't help matters), it will preserve some if not most of the rhubarb pie's tastiness, and indeed after some time the pie will be back on track towards ageing tastefully.[108]

[106]To repeat: the worthwhileness of an outcome does not go on to swallow up the axiological badness of the outcome (where this badness is understood only in terms of how good things are for persons). It might well be that the option of paying oblation to Moloch is ranked below the alternative on the remaining grounds.

[107]Gnon plays the role of catastrophic risk in this yarn.

[108]Paying oblation to Gnon in this case represents safety-engineering our world for the eventuality of a freak catastrophe; such that, if the population were to survive, then their lifetime welfare level

So long as we do not go extinct, and if there is the possibility of convalescing from whatever hellish state that Moloch placed us in, then the terrible thing of creating lives not worth living for some stretch of time, while neither better or worse for these poor creatures, will be part of a worthwhile attempt to achieve the best things in life. This means that (2) is false.

But if we do not recover, then this will not only leave the claims against paying Gnon fully intact, dragging down the overall goodness of the lottery, but the fact that persons would be badly off would further drag down overall goodness.

*
* *

There is no obvious answer to what humankind should now do with respect to Moloch and Gnon. Sure, placating Moloch, for instance, could be worthwhile. But it may not be. Humankind could get snuffed out by a freak bubble of true vacuum before reaching Shangri-La!

More so, as is implied by my statement in §4.4.1.6—specifically, 'I suppose that we have only scratched the surface of the best things possible in human history'—it is possible that both options being compared contain some of the best things in life.

My toy model is designed specifically for the purpose of resolving the first problem. I will generate a large sample space of outcomes that are representative of the many different possible human histories. Of course, this cannot tell us when to expect the best things in life, their magnitude, or prevalance. However, I think this doesn't present as a problem so long as we assume (rather safely, I think) that the more we mature (in terms of our development) the greater in both quantity and quality we would find these best things in life. (Note: I can think of no solid reason to believe this isn't true true no matter if we develop either dangerously quick or safe-n-slow.) This then gives us a rough idea of the expected loss in the best things in life to work with.

Building on this assumption, we can then resolve the second problem. There will likely be far more of the best things in life overall if we placate Moloch, given how much longer humankind might be around. We could then reduce the strength of persons' claims against placating Moloch down by the ratio of what's at stake (in expectation) between $\mathbb{L}_1$ and $\mathbb{L}_2$. If there would be, for example, twice as many of the best things in life in $\mathbb{L}_2$, then the claims against $\mathbb{L}_2$ would be reduced by some function describing this fact.

There will be many different ways to describe this function. For instance, it could drop the strength of competing claims at an exponentially growing rate as the ratio of how many of the best things in life were at stake grew larger. I'm going to leave open which of these is right. I am only vehemently opposed to reducing competing claims against the lottery with fewer best things in life down to nothing. After all, both lotteries contain the best things in life, and my intuition is that, while it is surely more valuable to have more of the best things in life, their addition doesn't seem to swamp the value described by the other considerations in the same way. After all, the

wouldn't drop as low.

burdens shouldered are no longer worthwhile in the same sense; these burdens merely improve the chances of even more of the best things in life being realized. This is, I believe (but have no argument for), importantly different from worthwhile burdens that preserve the very existence of the best things in life.

As I argued in §4.2, the multiple considerations which strengthen a person's claim against an option have finite force (so, limited contribution towards overall value). Some of them are, many of us agree, far less valuable or morally important than how good an outcome is for persons (e.g., equal chance of benefit). By contrast, the value of worthwhileness has no such upper bound, and will eventually swamp as we take the limit of the size of the timeless population to infinity.

Therein lies the central move of my argument.

Lotteries comprising only outcomes where the population doesn't survive beyond Earth can contain at most some number x of the best things in life. This includes Tim Mulgan's suggestion of the *Unambitious Policy*. By contrast, those lotteries that involve humankind evolving past Earth and spreading out across multiple galaxies could describe histories where our populations grows to be several orders of magnitude larger than the 'unambitious population'. This means—given our earlier assumption—that they will contain several orders of magnitude more of the best things in life (in expectation). And this, in turn, means that the ratio of what's at stake will eventually reduce the strength of competing claims to zero.[109]

Of course, I denied the possibility of an infinitely large population earlier this chapter. Still, I left open, and it seems reasonable to think, given what we know of physical eschatology, that the size of the proper subset of the timeless population which inhabits the post-Earth era could be astronomically huge. So, while this will never quite totally swamp the other considerations that make up the Competing Claims View, it's possible that, under certain functions describing the comparative weight of worthwhileness in the larger population, the strength of these competing claims becomes *very tiny* (but not vanishingly small, negligible)—this is to say, small enough that they should make absolutely no difference to substantive matters.

Thereby, in large population decisions of the sort that are of interest to us here, the strength of competing claims against (e.g.) placating Moloch or paying oblation to Gnon will be just about totally reduced to nothing. Therefore, we find that the Competing Claims View's evaluations of overall value in large population cases just about totally reduces back down to evaluations of what is good for persons. Therefore, the Competing Claims View generates approximately the same *cardinal ranking* of lotteries as a form of Averagism resulting from the combination of the VoIP and Conditionalism. And that is what I had set out to show in this chapter.

[109] We will write n to describe the number of 'copies' of the population that ever lives on Earth, E, that could be brought into existence if we colonized space. In the case of nE, as $|n| \to \infty$ we find that the corresponding number of copies of x goes up by approximately as much (if not slightly more given our descendants will (a) be more mature and capable of creating the best things in life; and (b) have greater access in the cosmos.)

4.6 Concluding Remarks

There are roughly four categories under which every outcome on my toy model falls:

(a) *go fast and suffer a broken world*;

(b) *go fast and early extinction*;

(c) *go safe-n-slow and early extinction; or*

(d) *go safe-n-slow and prolong humanity's place among the stars (with the occasional freak catastrophe temporarily lowering their lifetime welfare).*

I have argued that VA ranks (a) as worse than (b), and (c) as worse than (d). The Competing Claims View too shares this ranking as we have seen.

But I could not settle the matter of whether going slow was better on balance than going fast in this chapter. Indeed, whether (b) is better or worse than (d) itself is up for debate. The fact is there are infinitely many ways things could pan out either way we go—this is because there are too many random things that will also influence the outcome. No one, short of a fortune-teller with a very, very special crystal-ball, can know with certainty which outcome will obtain *this time around* if we decide (e.g.) to go slow. I don't see how this hullabaloo could be resolved from the armchair. As I have repeatedly said, my plan is to run the numbers on `Python`, such that we get a robust sample of how the details tend to get coloured in upon adopting some policy. It's about as close as we are ever going to get in terms of doing fieldwork in population ethics. With this in hand, we can then pick the policy (lottery) which is best according to the Competing Claims View (VA).

But before we run the toy model, I'm going to have to say a little bit more about the adoption of the go dangerously fast policy. Specifically, I need to address how dangerous we should be prepared to let things get when developing fast.

Part II

The Toy Model

5

How Fast is Too Fast?

> "It's always useful to know where a friend-and-relation is, whether you want him or whether you don't."
>
> A. A. Milne, *Pooh's Little Instruction Book*

5.1 Preliminaries

In the best case scenario, our population would go about creating 𝔼s as fast as possible, and they would get every step along the way right. Cumulative catastrophic risk would be very low or zero, and would never grow (no matter what they do). The go-fast policy, in a nutshell, proceeds as if they were in this felicitous state of affairs.

Of course, in reality they aren't so lucky. Their world is plagued by man-made risk. By going fast they threaten to impose the terrible burdens associated with such risks on subsequent members of the population (as well as themselves!). We must imagine persons adhering to such a policy, not as moral monsters, but instead as (decidedly) giving the upper-hand to those currently alive when acting. This is not insane. Nor is it even morally irresponsible on some population axiologies. Indeed, there are extant moral theories in the field of population ethics that prescribe exactly this much.

Last chapter I argued that doing what is best in terms of personal goodness amounts to the same thing as the fairest distribution of burdens, given that some rot or waste is inescapable. It is possible, I have said throughout, that the go-fast policy is best all-things-considered. But this was largely chewing the fat without making explicit what this kind of lottery entails. The purpose of this chapter is to make this plain.

If our forebears had followed the go-fast policy, then they would have thrown caution in the wind for the most part. Chiefly, they would not have taken on any burdens attributable to prolonging mankind as this would have lowered their own lifetime welfare (with reference to their basket of 𝔼s). This is a bad thing according to go-fast. The reverse—where they benefit at the cost of shortening mankind's lifespan—is a good thing. But since these 𝔼s are permanent features of the world (in the absence of

an existential catastrophe) once brought into existence, those persons brought about after their creation are able to benefit from them. So, there is a tradeoff here. Going faster means cutting humanity's lifespan short by some amount (how much depends in big part on humanity's starting conditions), but it also means that those that come into existence are better off than if their forebears had instead costly steps to reduce existential risk.

On the go-fast policy, increasing the risk of extinction is a neutral matter. This is because it's not worse for merely possible persons. Plus, since none of our forebears shouldered the burdens of placating Moloch, my first batch of arguments from §4.5.2 do not kick in.[1] What matters is improving the conditions faced by those already alive. But, this being said, the policy tells us that it would be a terrible thing if cumulative catastrophic risk culminated in a broken world where life is short, brutish, and nasty—in other words, if life were worth not living. More precisely, if these horrible lives were mere additions, in no way affecting the welfare of the existing (sub)population, then this should be avoided. It is, after all, a bad thing that contributes negatively towards our initial assessment of outcome goodness. But if this dangerous action would greatly improve the welfare of the existing population, then there would be a trade-off. It might well be the case that they ought to push the button and break the world if the rewards gained by them outweigh the harm done by enough according to the go-fast policy.[2] That's basically as much of a safety net as go-fast allows.

As noted above, various views in population ethics are adjoined by the better-worn intuitions responsible for these two features of the go-fast policy. But these intuitions have time and again proven themselves rather difficult to jointly pin down. Something always seems to go wrong somewhere along the way. We can name them, respectively, (a) the Non-Greedy Principle of Personal Good and (b) Prohibition on Miserable Mere Addition. Putting them together forms what I will hereafter call, after its leading proponent, *Broome's Intuition About Neutrality* ('$\mathbb{BN}$'). Many have given up on this project, given these arguably insurmountable problems, including John Broome himself. The purpose of the chapter is to demonstrate that the war has, in fact, not been lost; there is at least one combination of a (mathematically) well-behaved axiology and bridge principle that yields a moral theory which satisfies the normative reading of $\mathbb{BN}$. Armed with this moral theory, we will be able to pull out a more precise understanding of just how dangerous humanity is prepared to let things get under go-fast.

Bear in mind, although we are solving the problem in the familiar context where one would find this hullabaloo in the moral literature, the 'ought' belonging to the go-fast policy is not a moral ought. Its prescriptions are cashed out purely in terms of

[1] Bear in mind, I said that: 'We can make the world go better by prolonging humankind *but only if previous persons paid oblation'.*

[2] We can easily predict that a version of go-fast which was purely indifferent to the plight of subsequent persons would lose to go-slow on my toy model. The reason that this formulation of go-fast does not predictably lose to go-slow is that we can imagine possible worlds where (a) catastrophic risk is initially very low; (b) climbs up very slowly; and (c) they go extinct right before humanity would have plummeted into Hell. If that were so, then this would result in an average lifetime welfare level which outstrips that provided by go-slow under unfortunate starting conditions.

the personal value of ϕ-ing—this is to say, it has the form: 'if ϕ would be better for me than it is worse for any subsequent persons, then this provides me with (unchallenged) prudential reason to perform ϕ'. The 'moral ought' takes place at the level of evaluating this policy against its alternative. Still, it would be cumbersome to translate the moral arguments and theories into the appropriate context as we went along. So, this is something I postpone doing until the concluding section, in place of the ordinary practice of summarizing one's arguments or findings.

The present chapter requires being awake to some new notation and definitions. Let's start by unpacking the new, as well as rehearsing some of the old.

5.2 Basic Structure

This section of the chapters describes the basic structure of the framework I adopt, some light formal notation, and unpacks important background for us to bear in mind.[3]

Let the symbol $\mathcal{L}$ stand for the power set of all possible persons. The set is countably infinite; we are going to number its elements in some arbitrary order: {1, 2, ..., n}. Now we can define 'population'. My use of the term will be the same as John Broome's. I quote him in full:

> By the 'population' of a distribution, I mean the collection of all the individual people who, at some time or other, are alive in that distribution. I do not mean the *number* of those people; I shall call that the 'size' or the 'number' of the population. So in my terminology, two different populations may have the same size.
>
> A population is eternal: it consists of all the people who are ever alive.[4]

Different populations can be described by letters; e.g., $A = \{1, 2, 3\}$ and $B = \{4, 6\}$.

We will write B for the outcome describing the welfare distribution corresponding with population B. We will assume that there is a set of (lifetime) wellbeing levels equipped with an ordering $\geq$. I'll adopt the convention of writing B' to describe the outcome where the size and composition of the population remains the same and every member of B has a higher lifetime welfare level in B' than he does in B.

Welfare distributions are described by vectors. The symbol b_i is *either* a real number that stands for the lifetime welfare level of a person i that exists in an outcome B *or* an arbitrary non-numerical value, Ω, which holds the places of people that do not exist in B.[5] For example,

[3] Much of this framework is based on that supplied in (Broome, 2004, 25-26) and Cusbert and Kath (forthcoming).

[4] (Broome, 2004, 18)

[5] The field is split (unevenly, I think) on how to describe a person's 'null life' in those outcomes in which he does not exist. Some place a zero here instead (or have Ω stand for zero in order to avoid confusion when reading the vectors), while others maintain that this life has no welfare, not zero welfare.

$$\mathrm{A} = (a_1, a_2, a_3, a_4, ..., a_n) = (8, 4, 6, \Omega, ..., \Omega).$$

For sanity's sake, the three dots in A describe the remaining elements of the infinite-length vector. Say that the combination of A and B, written AB, is the union containing precisely the populations of A and B with the lifetime welfare levels described by those respective outcomes.[6] E.g., if A = (Ω, 2, ..., Ω) and B = (8, Ω, ..., 4), then AB = (8, 2, ..., 4).[7] The set of all vectors can be written as $\mathbb{A}$.

A population axiology is the set of principles which determines the value of a population. It encodes the betterness relations among outcomes, and thereby settles the betterness ordering of states of affairs. This ordering is expressed by the relations 'is better than' and 'is the same as'. These relations can be each defined in terms of a single primitive relation, 'at least as good', as follows (where '$\succ$', '$\sim$', and '$\succeq$' stand for respectively 'better', 'the same as', and 'at least as good'):

- $\mathrm{A} \succ \mathrm{B} \iff (\mathrm{A} \succeq \mathrm{B}$ and $\mathrm{B} \not\succeq \mathrm{A})$
- $\mathrm{A} \sim \mathrm{B} \iff (\mathrm{A} \succeq \mathrm{B}$ and $\mathrm{B} \succeq \mathrm{A})$

This primitive relation, however, does not apply in all cases. Two outcomes, A and B, might be incommensurable.[8]

We make the simplifying assumption that the threshold for a life worth living is 0.[9] Bearing that in mind, we can now add that a *happy outcome* is one containing no

[6](Cusbert and Kath, forthcoming, 4)

[7]Following Cusbert and Kath, we will say that if the populations of two outcomes overlap, then their combination is not well-defined.

[8]To be sure, incommensurability is not the same thing as vagueness. It isn't that there is no obvious answer to whether some outcome A is better than another B; rather, there is: *neither A nor B is either better than or equal to the other*. See (Broome, 2004, 169).

[9]This does not on its own suggest that it is better for the person that he lives a positive life than that he doesn't. On the one hand, it is open to us to adopt *Existence Comparativism*—meaning that, possible states of affairs in which p exists can be better or worse (or equally good) for p than possible states of affairs in which p doesn't exist (Pummer, forthcoming, 4). Notable proponents include: Fleurbaey and Voorhoeve (2015); Pummer (forthcoming); and Cusbert and Greaves (forthcoming). See also (Arrhenius and Rabinowicz, 2015, 429) for an account of *Limited Comparativism*: it can be better (worse) for someone to live than never live if he exists in the actual state of affairs, but it cannot be better (worse) for him if he doesn't—(cf. Holtug, 2001). Others have balked at the suggestion of these two states of affairs being comparable in any meaningful sense, going so far as to call it absurd—see especially (Broome, 1999, 168); Parfit (1984); and Bykvist (2007b). Important to bear in mind here is that even if one adopts Existence Noncomparativism, so long as he doesn't couple it with a *Strong Person-Affecting View*—according to which, a possible state of affairs is better (worse) than another only if it is better (worse) for someone—they can nevertheless claim that an outcome is worse even if it is not worse for anyone. It might be, for example, worse insofar as it is bad for (without being worse for) some persons. This is Broome's position. Tentatively, we will follow in Broome's footsteps and assume the truth of *Existence Noncomparativism*. After all, part of what we are trying to do here is see if we can derive a moral theory which, not just satisfies the normative reading of $\mathbb{BN}$, but on terms close-enough to what Broome himself would deem adequate. Every step we find ourselves taking away from his original goal is a cost for us to bear in mind.

person at a welfare level below 0.[10] Note that the empty outcome, in which no persons exist, is a happy outcome. Let $\mathcal{H}$ be the set of happy outcomes.

The structure of goodness all-things-considered can be seen as an aggregate (or balancing) of goodness with reference to a (or several) moral ideal(s) m. There might well be moral ideals, such as equality, that one wishes to incorporate into his axiological framework.[11] In this chapter, however, we are going to assume *welfarism*; the morally relevant features of an outcome are fully determined by its distribution of lifetime welfare.

Define an axiology on $\mathbb{A}$ as a set $\mathcal{R}$ of relevant respects such that for each $r \in \mathcal{R}$ there is: (a) a set of outcomes $\mathbb{A}_r \subseteq \mathbb{A}$ called the *outcomes that register on* r; and (b) a reflexive, transitive, binary relation $\succeq_r$ on $\mathbb{A}_r$.[12] We will assign to an outcome A a numerical score, r(A), describing how good A is with reference to some r. To say that some outcome A is r-better than B, we have: A $\succeq_r$B if and only if $r(\text{A}) > r(\text{B})$.[13] Let's also say that A and B are *r-incomparable* just in case neither A nor B is r-better than, r-worse than, or equally as r-good as the other.[14] Furthermore, we *aren't* going to assume that all outcomes register on any given respect (that is, $\mathbb{A}_r = \mathbb{A}$ for $\forall r \in \mathcal{R}$). If an outcome A doesn't register with reference to some r, then A is *r-incomparable* to all outcomes. To draw an analogy, it's akin to a category error to claim that the colour yellow is heavier than this teacup; the teacup has a well-defined weight unlike the colour yellow.

Define a decision as a finite set of outcomes $\mathcal{D} \subseteq \mathbb{A}$.[15] Where A $\in \mathcal{D}$, say that A is *available in* $\mathcal{D}$.[16] $\mathcal{D}$ contains precisely those outcomes among which the deciding agent must choose.

Something must now be said about how axiology relates to a moral theory, $\mathcal{M}$, which delineates permissible actions—this is to say, the subset $\mathcal{M}(\mathcal{D})$ of $\mathcal{D}$. We will require a *bridge principle*, ζ, as a function that takes axiology and any decision $\mathcal{D}$ as input, returning a subset $\zeta(\mathcal{R}, \mathcal{D})$ of $\mathcal{D}$.[17] These are the permissible acts in $\mathcal{D}$ according to both $\mathcal{R}$ and ζ. It can be said of $\mathcal{M}$ that it cuts mustard if and only if the permissible outcomes according to $\zeta(\mathcal{R}, \mathcal{D})$ are precisely those of $\mathcal{M}(\mathcal{D})$.

Of course, betterness-facts do not fully settle the ought-facts on some moral theories. Some moral theories might even wildly slew from the theory of goodness underpinning them when governing either what we ought to do or what we owe each other.[18] I do

[10](Cusbert and Kath, forthcoming, 4)

[11]E.g., Temkin (1993).

[12](Cusbert and Kath, forthcoming, 4)

[13]In order to know *by how much* one outcome is better compared to another, we can follow (Cusbert and Kath, forthcoming, 5) in substituting (b) for (b*): a two-placed, integer-valued *difference function* δ: $\mathbb{A}_r \times \mathbb{A}_r \longrightarrow \mathbb{Z}$ such that for all A, B, C $\in \mathbb{A}_r$: (i) δ_r(A, A) = 0; and (ii) δ_r(A, C) = δ_r(A, B) + δ_r(B, C). We are going to read δ_r(A, B) = x as stating that A is x units r-better than B. This may be abbreviated to (A $\succ_r^x$ B).

[14](Cusbert and Kath, forthcoming, 4)

[15](Cusbert and Kath, forthcoming, 6)

[16](Cusbert and Kath, forthcoming, 6)

[17](Cusbert and Kath, forthcoming, 6)

[18]But not too much. As John Rawls once forcefully put it: "All ethical doctrines worth our attention

not have the space to explore these kinds of moral theories in the paper. Instead, I will confine the search for a class of moral theories to those constituted by a bridge principle which, as a bare minimum, satisfies:

> (†): For every axiology $\mathcal{R}$ and bridge principle ζ, if in decision {A, B}, A is permissible and B is impermissible according to $\mathcal{R}$ and ζ, then there exists $r \in \mathcal{R}$ such that A $\succ_r$ B.[19]

This constraint tells us that in any two-outcome decision, an outcome can be impermissible only if it is worse in *some morally relevant respect* than the other outcome.[20]

Finally, below are four principles that we will occasionally refer to in subsequent sections of the paper.

- *Principle of Personal Good.* Suppose two distributions have the same population of people. If both outcomes are equally good for each member of the population, then these two outcomes are equally good. Also, if the first outcome is at least as good as the second outcome for each member of the population, and if the first outcome is better than the second for some member of the population, then the first outcome is better than the second outcome.[21]

- *Impartiality Between People.* If two one-dimensional distributions are described by vectors that are permutations of each other, the distributions are equally good.[22]

- *Principle of Equal Existence.* Suppose two distributions have the same population of people, except that an extra person exists in one who does not exist in the other. Suppose each person who exists in both distributions is equally well off in one as she is in the other. Then there is some range of wellbeings (called the 'neutral range') such that, if the extra person's wellbeing is within this range, the two distributions are equally good.[23]

- *Principle of Incommensurate Existence.* Suppose two distributions have the same population of people, except that an extra person exists in one who does not exist in the other. Suppose each person who exists in both distributions is equally as well off in one as she is in the other. Then there is some range of wellbeings (called the 'neutral range') such that, if the extra person's wellbeing is within this range, the two distributions are incommensurate in value.[24]

take consequences into account in judging rightness. One which did not would simply be irrational, crazy"" (Rawls, 1971, 30).

[19](Cusbert and Kath, forthcoming, 7)

[20]If it weren't, then the impermissibility of B would have to be derived from something totally alien to B's goodness relative to A—and that's, well, mighty spooky for some of us. (Bear in mind that the category of 'spooked' will include those of us subscribing to non-teleological theories that adopt something like T. M. Scanlon's *Buck-Passing Account* Scanlon (1998); (cf. Frick, 2014).)

[21](Broome, 2004, 120)

[22](Broome, 2004, 135)

[23](Broome, 2004, 146)

[24](Broome, 2004, 167)

5.3 Broome's Intuition About Neutrality

One of the better-worn intuitions in the field of population ethics is that adding a person to the world is morally neutral in terms of our theory of goodness.[25] Broome puts it the following way:

> [The] presence of an extra person in the world is neither good nor bad. More precisely: a world that contains an extra person is neither better nor worse than a world that does not contain her but is the same in other respects.[26]

Three comments. First, it's worth repeating that we are talking only about axiological (or evaluative) neutrality.[27] Second, for the purpose of our discussion, the range of lifetime welfare levels which are neutral is understood to have a lower bound, but no upper bound. This is to say, the addition of a person that would have a miserable life—one in which he would kill himself if only he could—doesn't have neutral (contributive) value; the world goes worse, all else being equal, if he is brought into existence. Let's call this the act of *miserable mere addition*.

My third comment is most crucial. Broome's claim is only that we are morally neutral about mere additions. But let us imagine that bringing a person about would have some influence over our own lifetime welfare.

The intitial description of the Intuition of Neutrality provided by Broome leaves open whether, if we are adhering by the Principle of Incommensurate Existence, the person has *some* effect on the world even though his "addition has no positive or negative value in itself".[28] Picking up on this ambiguity, Włodek Rabinowicz has written,

> However, if an extra person is added to the world, the value of the world does *not* remain the same. While the result of addition is neither better nor worse, it is not equally as good either. It is incommensurate. Thus, additions of extra persons are not axiologically neutral in [the sense that they make *no change* to the value of the world].[29]

There's no obvious reason to ignore this change to the value of the world, looking only to how the original population would be affected, when determining outcome goodness. The big suggestion being that adding a happy person to the world might have a value that can be set against the value of other things in some other way.

[25](Broome, 2004, 143)

[26](Broome, 2005, 401)

[27]Recognize too that one need not maintain that this life would be neutral for the person living it. It might be well worth living in terms of personal goodness. If so, then this means we will have given up what Johan Gustafsson calls the *Equivalence of Personal and Contributive Value* (Gustafsson, forthcoming, 8).

[28](Broome, 2004, 146)

[29](Rabinowicz, 2009a, 399); cf. (Broome, 2009)

Broome rejects this.[30] In cases of this sort we should instead be guided by the ol' adage that what matters is making persons happy, not happy persons.[31] As Broome explains, a feature of the intuitive idea of neutrality is that if two things happen together, where one is bad, and the other neutral, then "[intuitively], the net effect of the two things should be bad. A bad thing combined with a neutral thing should be bad".[32] Placed in the context of variable populations, he maintains that adding a happy person to the world should not have a value that can be set against the value of other things in some other way. On this formulation of neutrality, malign addition makes the world go worse, and benign addition makes the world go better.

The Principle of Personal Good advocated by Broome can now be modified so that it handles variable populations involving additions whose lifetime welfare levels lie in the neutral range.

> *Non-Greedy Principle of Personal Good.* Take two distributions A and AB. If the extra persons contained in B all have welfare levels which lie within the neutral range, then if AB is all-things-considered worse (better) for the members of A than AB, then AB as a whole is worse (better) than A, and if AB is all-things-considered equally good for the members of A, then AB as a whole is neither better than nor worse than A.[33]

You may have realized that the Non-Greedy Principle of Personal Good incorporates the disjunction of Broome's Principle of Equal Existence and his Principle of Incommensurate Existence. This is how it is intended. My arguments do not turn on whether we interpret neutrality as being equally good as or incommensurable.[34]

[30] He presents his Argument from Greediness at (Broome, 2004, 169-170). The take-away message being: "Incommensurateness is not neutrality as it intuitively should be. It is a sort of greedy neutrality, which is capable of swallowing up badness or goodness and neutralizing it" (Broome, 2004, 170).

[31] (Narveson, 1973)

[32] (Broome, 2004, 169)

[33] This is a slightly repolished version of Frick's restatement of Broome's claims concerning the 'intuitive feature of neutrality' (Frick, 2017, 354). This principle can be understood four different ways when there are several options in a given decision. To illustrate this, suppose we are confronted with a three-option decision: {A', AB, AB'}. Firstly, we can take either an inclusive or exclusive stance on whether the loss suffered by $b \in$ B if outcome AB were to obtain should get set against the loss suffered by $a \in$ A in outcomes AB or AB'. The exclusive stance tells us that the comparative harm to contingent persons *should not* get set against that of necessary persons—this is to say, perturbations to the lifetime welfare of $a \in$ A ought to settle the matter in cases of benign or malign addition. I understand $\mathbb{BN}$ to take this hard line. Secondly, the Non-Greedy Principle of Personal Good can be understood in a stronger or weaker sense depending on whether the bad thing of creating a miserable life not worth living can be neutralized by bringing about a life worth living as well. My initial response is to adopt the stronger interpretation in which the combination of a bad thing (bringing a miserable person into existence) and a neutral thing (bringing a very happy person into existence) results in a bad thing all-things-considered. However, I go on to consider the alternative later in the chapter.

[34] The main difference between neutrality as equally good and neutrality as incommensurateness is that, if two things are equally good, any improvement (deterioration) of one of them would make it better (worse) than the other. By contrast, if two things are incomparable, then some improvement (deterioration) of one of the things won't make the thing better (worse) than the other thing. Joseph Raz has dubbed this the 'mark of incommensurability' (Raz, 1986, 325-326).

The conjunction of the Non-Greedy Principle of Personal Good and a prohibition on miserable mere addition results in what we will refer to as *Broome's Intuition About Neutrality* (hereafter abbreviated '$\mathbb{BN}$'). More precisely, it states: (a) mere addition is neutral in terms of outcome goodness; (b) mere miserable additions make the world go worse; and (c) benign (malign) addition makes the world go better (worse).

5.4 The Hunt Begins...

The aim of this chapter is to find a well-behaved axiology and bridge principle that, in combination, yield a moral theory that satisfies the *normative reading* of $\mathbb{BN}$.[35]

Without making a mountain out of a molehill, managing this is going be more difficult than may seem at first blush. There are many ways things could go wrong when we attach a bridge principle, even if our axiological framework is coherent. After all, the bridge principle that gets attached must pump out moral propositions consistent with our axiology for every $\mathcal{D} \subseteq \mathbb{A}$. This means that if we are neutral between every outcome in our decision, then all outcomes must be permissible. If one is better than all the rest, then only this outcome could be permissible. And so on. The difficulty of the task at hand will, at any rate, become apparent soon enough.

Below, a first step in the right direction is taken by abandoning an Overall Betterness Framework for a Multidimensional Framework.

5.4.1 Overall Betterness Framework

As Broome himself argues, the $\mathbb{BN}$ is untenable within an Overall Betterness Framework because it is internally inconsistent, so misbehaved.

[35]There are some moral theories which, though built atop an axiology which is a total mess, end up producing the intended propositions when partnered with a particular bridge principle. For example, Daniel Cohen purports to show that Moral Actualism—sometimes called Exlusion (Roberts, 2010b) or Axiological Actualism (Parsons, 2002)—produces a moral theory in the ballpark of $\mathbb{BN}$ when thought of the right way (Cohen, forthcoming). Moral Actualism is marred by misbehaviour, though. For starters, it violates *Axiological Invariance*: This is the requirement that the value of a state of affairs is independent of which state of affairs is actual. See especially (Broome, 2004, 74). The normative flavour of this plausible constraint is provided in (Carlson, 1995, 100), which Carlson attributes to Włodek Rabinowicz. Second, even if that somehow weren't deeply troubling, it cannot satisfy another plausible constraint which Krister Bykvist describes:

> *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, 2007a, 116).

As Cohen's argument makes clear, his overall moral theory only suffers from the first of these two problems, and (arguably) supplies the right normative claims despite it. Nonetheless, I don't plan to explore moral theories of this kind in the paper. My goal is, so to speak, to have my cake and eat it too. This means demonstrating that there is a class of moral theories consistent with $\mathbb{BN}$ at both the evaluative and normative level. (It's worth sharing however that, given Melinda Roberts' theory and Cohen's theory are extensionally equivalent with reference to their normative output, his theory is just as vulnerable to my counterarguments to Variabilism below.)

5. HOW FAST IS TOO FAST?

Let's begin by unpacking the Overall Betterness Framework. It describes there being a relation of overall betterness among outcomes, which alone determines permissibility in decisions.[36] An axiology is neutral between outcomes B and BG if and only if it finds that neither B nor BG is better overall than the other. If we were going to achieve the goal of the paper, the overall betterness relation would have to satisfy the following four conditions:

> *Overall Axiology* [O]. Let A, B, and H be populations such that $H \in \mathcal{H}$. Let $U \notin \mathcal{H}$. Let B' > B. Then:
>
> [O1] $A \succ AU$
>
> [O2] $(A \not\succ AH) \wedge (A \not\prec AH)$
>
> [O3] $AB'H \succ AB$
>
> [O4] $ABH \prec AB'$

Put informally, this reads: [O1] refraining from causing a miserable mere addition to exist makes the world go better; [O2] the mere addition of a happy person makes the world go neither better nor worse; [O3] the benign addition of a happy person makes the world go better;[37] and [O4] the malign addition of a happy person makes the world go worse.[38]

To get some normative meat on the bone, we can then apply a standard maximizing consequentialist bridge principle [C] in order to derive the subset $\zeta(\mathcal{R}, \mathcal{D})$ of $\mathcal{D}$.

> [C]. For all decisions $\mathcal{D}$ and all $A \in \mathcal{D}$, A is permissible in $\mathcal{D}$ iff $A \not\prec B$ for all $B \in \mathcal{D}$.[39]

Cusbert and Kath have dubbed the conjunction of [O] and [C] the *Overall Goodness Account* [OC]. We will call it this as well.

Regrettably, this framework is plagued with internal inconsistency. This can be shown with one of Broome's own toy examples:[40]

$$\begin{gathered} AB' = (4, 4, ..., 4, 6, \Omega), \\ AB'C = (4, 4, ..., 4, 6, 1), \\ ABC' = (4, 4, ..., 4, 4, 4), \\ AB = (4, 4, ..., 4, 4, \Omega). \end{gathered}$$

[36] (Cusbert and Kath, forthcoming, 8)

[37] For the moment, we stay silent on whether the benign addition of a miserable person makes the world go better, worse, or the same.

[38] As Cusbert and Kath have demonstrated, 'mere benevolence' falls out of [O3]. If A and H are taken to be empty outcomes, then [O2] entails that if B' > B, then B' makes the world go better.

[39] (Cusbert and Kath, forthcoming, 9)

[40] There are some other problems too. See (Thomas, 2016, 203ff).

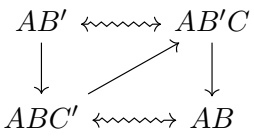


Figure 5.1: When reading the diagrams, a solid arrow from an outcome A to another B represents A being better than B. Its squiggly-counterpart describes its negation. For example, a double-headed squiggly arrow indicates that neither A nor B are better than the other.

In this decision, Ω is an arbitrary non-numerical value which represents non-existence.[41]

On Broome's suggestion, let's suppose that ABC' is better than AB'C. Although B is two units worse off in terms of his welfare, C is three units better off. (Plus members in distribution ABC' are equally well off.) This is plausible; however, it now places us in a very strange position for comparing the relative value of AB' and ABC'. There are three options available to us. (a) AB' could be better than ABC'. (b) AB' could be worse than ABC'. Finally, (c) AB' could be neither better than nor worse than ABC'.

The Non-Greedy Principle of Personal Good tells us that only (a) could be true.[42] But could ABC' be worse than AB? It could not. Suppose, for *reductio*, that ABC' were worse than AB'. If so, then AB'C would be worse AB' (by transitivity). But AB'C is not worse than AB'—it is neither better than nor worse than it (by [O2]). Therefore, AB' cannot be better than ABC'.[43]

[O] is internally inconsistent. It tells us at one and the same time that AB' is better than ABC' and that AB' is not better than ABC'. See Figure 1. So, to be sure, [OC]

[41]The field is split (unevenly, I think) on how to describe a person's 'null life' in those outcomes in which he does not exist. Some place a zero here instead (or have Ω stand for zero in order to avoid confusion when reading the vectors), while others maintain that this life has no welfare, not zero welfare. On the former, it is open to us to adopt Existence Comparativism—meaning that, possible states of affairs in which p exists can be better or worse (or equally good) for p than possible states of affairs in which p doesn't exist (Pummer, forthcoming, 4). Notable proponents include: (Fleurbaey and Voorhoeve, 2015); (Pummer, forthcoming); and (Cusbert and Greaves, forthcoming). See also (Arrhenius and Rabinowicz, 2015, 429) for an account of *Limited Comparativism*: it can be better (worse) for someone to live than never live if he exists in the actual state of affairs, but it cannot be better (worse) for him if he doesn't (cf. (**?**)). Others have balked at the suggestion of these two states of affairs being comparable in any meaningful sense, going so far as to call it absurd—see especially (Broome, 1999, 168); (Parfit, 1984); and (Bykvist, 2007b). Important to bear in mind here is that even if one adopts Existence Noncomparativism, so long as he doesn't couple it with a *Strong Person-Affecting View*—according to which, a possible state of affairs is better (worse) than another only if it is better (worse) for someone—they can nevertheless claim that an outcome is worse even if it is not worse for anyone. It might be, for example, worse insofar as it is bad for (without being worse for) some persons. This is Broome's position. Tentatively, we will follow in Broome's footsteps and assume the truth of Existence Noncomparativism. After all, part of what we are trying to do here is see if we can derive a moral theory which, not just satisfies the normative reading of $\mathbb{BN}$, but on terms close-enough to what Broome himself would deem adequate. Every step we find ourselves taking away from his original goal is a cost for us to bear in mind.

[42]Specifically, [O4] dictates this much.

[43](Broome, 2004, 169-170)

too crumbles in consequence.

Since there is no obvious remedy for this problem, and no other interpretation of the Intuition of Neutrality suggests itself, we are left with no choice but to throw out $\mathbb{BN}$, and along with it the Intuition of Neutrality too.

Or so says Broome.

5.4.2 Multidimensional Betterness Framework

The toy example only demonstrates that $\mathbb{BN}$ suffers internal inconsistencies if our framework relies on an overall betterness relation. We could drop this dependence.

Indeed, this is just what John Cusbert and Robyn Kath have gone on to do.[44] Their arguments show that if there is more than one respect in which an outcome can be better (or worse) than another, then both the permissibility of mere addition and the Non-Greedy Principle of Personal Good can be satisfied by a coherent moral theory. However, missing from their account is a moral prohibition on miserable mere additions. So, this is something that we will have to build in ourselves.

Let's say that A is better than B according to an axiology if it states that A is better than B is *some* relevant respect, and worse than B in *no* relevant respect. Similarly, if neither A nor B is better than the other in any relevant respect, then we will say that an axiology is neutral between A and B.[45] Padding out Cusbert and Kath's proposed axiology, we have the following alternative to the overall betterness framework:[46]

> *Respects Axiology* [R]. Let A, B, and H be populations such that $\mathrm{H} \in \mathcal{H}$. Let $\mathrm{U} \notin \mathcal{H}$. Let B' > B. Then:
>
> [R1] $(\exists r \in \mathcal{R})(\mathrm{A} \succ_r \mathrm{AU}) \wedge (\forall r \in \mathcal{R})(\mathrm{A} \not\prec_r \mathrm{AU})$
>
> [R2] $(\forall r \in \mathcal{R})(\mathrm{AH} \not\succ_r \mathrm{A}) \wedge (\forall r \in \mathcal{R})(\mathrm{AH} \not\prec_r \mathrm{A})$
>
> [R3] $(\exists r \in \mathcal{R})(\mathrm{AB'H} \succ_r \mathrm{AB}) \wedge (\forall r \in \mathcal{R})(\mathrm{AB'H} \not\prec_r \mathrm{AB})$
>
> [R4] $(\exists r \in \mathcal{R})(\mathrm{ABH} \prec_r \mathrm{AB'}) \wedge (\forall r \in \mathcal{R})(\mathrm{ABH} \not\succ_r \mathrm{AB'})$

Translation: [R1] miserable mere addition makes the world worse in one respect without making it better in any respect; [R2] mere happy addition makes the world go neither better nor worse in any respect; [R3] benign happy addition makes the world go better in one respect without making it worse in any respect; and [R4] malign happy addition makes the world go worse in one respect without making it better in any respect.

[R] is a partial axiological framework. It constrains the betterness relations among outcomes without fully describing them. For our purposes, we need to complete [R] so that it describes a betterness relation between every two-outcome comparison. Only

[44] See (Cusbert and Kath, forthcoming).

[45] It is now open to us as to how we interpret neutrality. We can adopt either the Principle of Equal Existence or the Principle of Incommensurate Existence. It's hard to see either one as having a clear advantage over the other. So I'll continue making use of their disjunction. We are neutral about an extra person if their addition neither makes the world go better nor worse.

[46] (Cusbert and Kath, forthcoming, 12)

then can we attach a bridge principle to our axiological framework, and determine whether our moral theory is compatible with $\mathbb{BN}$. To this end, let's suppose that there is a morally relevant respect of betterness for every $p \in \mathcal{L}$, such that (A $\succ_p^x$ AB) if and only if p exists in both A and AB and the difference between p's lifetime welfare in A and his lifetime welfare in AB is x.[47] This implies that the set $\mathbb{A}_p$ of outcomes that register on p contains only those outcomes in which p exists. Every other outcome is p-incomparable.[48] Cusbert and Kath call this the *quantitative personal axiology*, [P].[49,50] We will too.

To be sure, there are undoubtedly some number of other ways we could have formed a complete axiology on the back of [R]. This is, at first blush, the most natural though. There's nothing weird about it. Most of us *do* subscribe to this, and the rest of us can at least understand how *betterness for an individual* could be morally relevant in determining outcome goodness.

Plus, [P] supplies an adequate response to one of Broome's three doubts about neutrality *qua* incommensurateness.[51] As he puts it,

> We are not dealing with differing values. One option has a different number of people from the other. Whatever the value of people might be, each option realizes that value; one simply realizes a greater quantity of it than the other. So if our options are really incommensurate in value, we need some explanation of why. (...) But it looks like a fudge unless we can offer some reason why the neutrality of existence really amounts to incommensurateness rather than equality. This account requires some more explaining; that is my first doubt about it.[52]

The explanation for this, according to [P], is that if we can bring about one of two persons, both of which will be moderately well off, but one more so than the other, then it cannot be better for the better-off of them that we bring him about instead of the other. These two outcomes are p-incomparable in terms of outcome goodness.

Letting [P] loose on our earlier toy example, we find that there is no respect in which either AB'C or AB' is better than the other (by [R2]). AB' is better than ABC' in some respect, $r1$, and worse in none (by [R4]). AB'C is better in some respect, $r1$, than AB, and worse in none (by [R3]). Finally, AB'C is better in some respect, $r1$, than ABC'; but ABC' is better in another respect, $r2$, than AB'C. See Figure 2: [P]-Solution to Broome's Toy Example. These claims are all consistent given that we understand $r1$ to refer to the welfare of B, and $r2$ refers to the welfare of C.

[47] (Cusbert and Kath, forthcoming, 5-6)

[48] Again, we are going to lift this assumption of Existence Noncomparativism at a later stage of the argument.

[49] (Cusbert and Kath, forthcoming, 5)

[50] Cusbert and Kath demonstrate that [P] satisfies a weaker version of [R]; specifically, a version that is missing [R1]. See (Cusbert and Kath, forthcoming, 13).

[51] His other two doubts are (a) greediness and (b) the incompatibility of incommensurateness and vagueness. We are in the process of indirectly tackling (a). But we will not dive into (b). For arguments on this topic see especially (Rabinowicz, 2009b); (Sugden, 2009); cf. (Mozaffar, 2012).

[52] (Broome, 2004, 168)

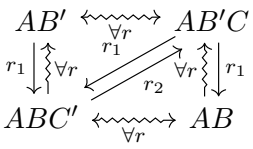


Figure 5.2: [P]-Solution to Broome's Toy Example.

So far so good. There are two crucial problems we must head off though. The first of these has been raised by Derek Parfit, as well as Broome. Consider the following decision:

$$\begin{aligned} \text{A} &= (1, \Omega), \\ \text{B} &= (\Omega, 1), \\ \text{B'} &= (\Omega, 2). \end{aligned}$$

According to the Principle of Personal Good, B' is better than B. Also, Impartiality Between People tells us that A and B are equally good. Therefore, B' must be better than A.[53]

However, B' is not better than A for either the first person or the second person. So, [P] tells us that B' is not better than A in any respect. This is best-known as the Nonidentity Problem[54], and it arises because [P] simply denies that A and B are equally good, even if they are mere permutations of each other. Bear in mind, they are *incommensurable*. Looking at our options, Broome maintains that we ought to give up [P] before giving up impartiality. He writes: "So much the worse for [P]. [B'] is indeed better than A. (...) When [P] says otherwise, it is wrong".[55]

There are three different counter-replies which could be offered up here. For starters, we could simply deny the claim that B' is better than A. To be sure, not everyone shares Broome's intuition, and the Nonidentity Problem is nowhere near being open and shut. Among those fueling the dissent, one will find Melinda Roberts who writes that "[we] can discern no (...) morally significant loss [in such cases]. On the other hand, neither do we clearly in such cases discern wrongdoing. (When [all else is equal], is it really

[53]This supposes that if A and B are at least as good as each other, then B' is better or worse than A if and only if it is correspondingly better or worse than B (Broome, 2004, 21). And *this* presupposes the truth of the *Independence of Irrelevant Alternatives*: how good an outcome is doesn't depend on which outcome it is being compared to, nor on what other outcomes are available. Option set independence—this is to say, contraction and expansion consistency—is implied by the *Internal Aspects View of Outcome Goodness* (Temkin, 2012, 370). For well-behaved axiologies (that do not involve incompleteness) this entails the *Principle of Like Comparability of Equivalents*: "if two outcomes or prospects are equivalent (meaning equally good) in some respect, then however the first of those outcomes or prospects compares to a third outcome or prospect in that respect, that is how the second of those outcomes or prospects compares to the third outcome or prospect in that respect" (Temkin, 2012, 237).

[54](Parfit, 1984, 351ff)

[55](Broome, 2004, 136); cf. (Broome, 2004, 145)

wrong to bring a genetically impaired but happy child into existence rather than a less impaired and happier child into existence?)"[56]

The issue raised by Roberts is no storm in a teacup. But one will only get so far relying on this line of defense. My gut feeling is that this is just one of those kinds of cases where no one is going to budge. The more promising approach is to demonstrate that a major element of $\mathbb{BN}$, the Non-Greedy Principle of Personal Good, is itself incompatible with Impartiality Between People. As Broome himself notes, a credible implication of [P] is that a person's coming into existence makes him neither better nor worse off than he would have been; so, it is morally neutral. While Broome thinks we should reject [P] for its other implications, the Non-Greedy Principle of Personal Good is one we can hold onto independently. But, as Johann Frick has argued, Broome's reason for rejecting [P] should lead him to also reject the Non-Greedy Principle of Personal Good. Here's how Frick's argument goes.[57]

Let's imagine the following decision:

$$\text{AG'} = (4, 4, ..., 2, 5, \Omega),$$
$$\text{AGH'} = (4, 4, ..., 2, 4, 4),$$
$$\text{A'GH} = (4, 4, ..., 4, 4, 2).$$

The Non-Greedy Principle of Personal Good implies that the move from AG' to AGH' must be a change for the worse since it is all-things-considered worse for G. The addition of H' can alter this conclusion only on pain of greediness. So, AG' is better than AGH'. Consider next the move from AG' to A'GH. Although G is a little worse off, A is left even better off. The Non-Greedy Principle of Personal Good therefore tells us that A'GH must be all-things-considered better than AG'. Therefore, by transitivity, A'GH must be better than AGH'. But these two outcomes are mere permutations of one another. According to Broome's principle, Impartiality Between People, A'GH must be equally good as AGH'. Therefore, the combination of the Non-Greedy Principle of Personal Good and Impartiality Between People generates a contradiction.

Thereby, it is not just [P] on the chopping block, but the Non-Greedy Principle of Personal Good too.

It is, of course, open to us to just throw our hands up and give up on $\mathbb{BN}$.[58] This would be an understandable reaction. Another is to accept that we *aren't* impartial about the matter. To repeat our mantra: *we are all about making persons happy, not making happy persons.* I won't argue that the second option is best. Frankly, I'm not convinced it is. Rather, my humble goal is merely to find out if a moral theory can satisfy $\mathbb{BN}$ at the normative level if our moral evaluations turn on how persons are made better or worse off. I can easily imagine some of us taking this approach and sticking to their guns that A is at least as good as B'. At any rate, this is not so obviously misguided a stance that we can dismiss it in true knee-jerk fashion.

[56](Roberts, 2010b, 362)
[57]The argument can be found at (Frick, 2017, 356-358).
[58]Indeed, this is precisely how Broome himself responds.

Our third option is owed to Teru Thomas, and seems to me to be as good as if not better than our second.[59] His suggestion is to use a counterpart relation between the persons in two different outcomes. The basic idea being that, relative to such a counterpart relation, persons with counterparts are to be fully accounted for when evaluating the final value of an outcome.

This relation must satisfy two constraints.[60] (a) It should extend the relation of transworld identity. If a person p exists in both outcomes, then he must be his own counterpart. (b) The counterpart relation must pair up as many contingent persons as possible from one outcome with contingent persons in another outcome. Because there are possibly very many counterpart relations meeting these two constraints, we need some kind of system for determining the final value of AB in a decision {A, AB}. For our purposes, we shall adopt the same model as Thomas—what he calls *Complex Necessitarianism* (hereafter abbreviated '$\mathbb{CN}$'). In order to make this precise, we will require three statistics:[61]

- $\mathcal{W}(\mathrm{A})$, the total welfare of the necessary people in {A, AB};
- $\mathcal{N}(B)$, the number of contingent people in {A, AB};
- $\mathcal{W}(\mathrm{B})$, their total welfare.

His model can now be formulated as follows.[62] Let the final value of A be

$$V(\mathrm{A}) = \mathcal{W}(\mathrm{A}).$$

This is because they are the necessary persons. The final value of AB is a bit more complicated:

$$V(\mathrm{AB}) = \mathcal{W}(\mathrm{A}) + X,$$

where X represents the value of B. It should just be the average of the value of B relative to all candidate counterpart relations. Importantly, there is one exception to this rule. If the contingent persons in B have lives not worth living, then this fully counts against B.[63]

$$X = \begin{cases} \mathcal{W}(\mathrm{B}), & \text{if } \mathcal{W}(\mathrm{B}) \leq 0. \\ \min(\mathcal{N}(\mathrm{A}), \mathcal{N}(B)) \frac{\mathcal{W}(\mathrm{B})}{\mathcal{N}(B)}, & \text{otherwise.} \end{cases}$$

[59] See (Thomas, 2016).
[60] (Thomas, 2016, 210-211)
[61] (Thomas, 2016, 211)
[62] (Thomas, 2016, 212)
[63] Thomas' formula is a little stronger than this as one may glean below. It also holds that a life which is of personal neutral value also fully counts towards the value of B. This won't feature in any of the toy examples subsequently considered.

Here's the payoff. If there are only two contingent persons in some decision, p_1 in A, and p_2 in B, then they are each other's counterpart in a decision {A, B, B'}. In this way we can claim, for example, that B' is (*de dicto*) better for the person that does get brought into existence than A, and no longer do we get gored on the Nonidentity Problem.

This is only the core idea behind Thomas' proposed moral theory. It would confuse matters to try to handle more problematic toy examples at this stage. This is to say, I must leave things in an awfully nebulous place for the moment so that we can introduce the other bits of machinery on which Thomas' fuller moral theory operates. To be sure, though, the above has not been a waste of either space or time, frivolous. $\mathbb{CN}$ plays a vital role in my subsequent discussion.

To sum up, we really just have two options on the table. We can accept that B' is not better than A; rather, they are incommensurate. Alternatively, we can adopt $\mathbb{CN}$ (or something very much like it). Moving forward, I will run on the assumption that both options are equally plausible.

*
* *

Before stating the second problem for [P], I'll describe the partial bridge principle put forward by Cusbert and Kath. This bridge principle will shortly after need to be fleshed out by us.

Note that we no longer have access to [C], given [C] is formulated in terms of an overall betterness relation. But we are not stuck twiddling our thumbs either. The following proposal from Cusbert and Kath is intuitively robust.[64] In order to properly understand it we must be awake to a new piece of terminology. For any decision $\mathcal{D}$ and any available outcome A $\in \mathcal{D}$, say that A is *impeccable* in $\mathcal{D}$ just in case A $\not\prec_r$ B for all $r \in \mathcal{R}$ and all B $\in \mathcal{D}$.[65] In other words, an outcome is impeccable whenever it's no worse in any respect than any available outcome. Bearing this in mind, we can add:[66]

> [B]. For all decisions $\mathcal{D}$ and all A, B $\in \mathcal{D}$:
>
> > [B1] If A is impeccable in $\mathcal{D}$, then A is permissible in $\mathcal{D}$.
> > [B2] If there exists A $\in \mathcal{D}$ and $r \in \mathcal{R}$ such that A is impeccable in $\mathcal{D}$ and B $\prec_r$ A, then B is impermissible in $\mathcal{D}$.

According to [B], impeccable outcomes are permissible, and an outcome is impermissible if it's worse in *any* respect than any available impeccable outcome.

[64] More so, as Cubert and Kath explain, it does a good job of mapping onto the same consequentialist concerns. For starters, it attributes (im)permissibility purely on the basis of the betterness relations holding among available outcomes. Second, it satisfies (†). Third, an outcome is deemed permissible provided that it is no worse than any of the alternatives. See (Cusbert and Kath, forthcoming, 11).

[65] (Cusbert and Kath, forthcoming, 12-13)

[66] (Cusbert and Kath, forthcoming, 13)

Equipped with an axiological framework and partial bridge principle, our problem is this. There's no obvious way to make heads or tails of what is meant by A $\succ_r$ AU in [R1]. After all, A and AU are u-incomparable.[67] Put differently, the same reason why [R2] is able to remain neutral between AH and A is what's preventing us now from breaking the silence regarding miserable mere additions.

Moreover, even if we could somehow solve this initial problem, there are going to be many decisions where no impeccable outcome obtains. In order to handle cases of this sort, we will need to put some meat on the bones of [B]. The obvious move here is to define $\zeta(\mathcal{R}, \mathcal{D})$ such that permissible outcomes are those with "the least to be said against them in the context of that decision".[68] This requires yet some more terminology. Define the *shortfall* of A in $\mathcal{D}$ as the sum of the amounts by which A is worse in every relevant respect than the outcome in $\mathcal{D}$ that does best in that respect.[69]

> [S]. For all decisions $\mathcal{D}$ and all A, B $\in \mathcal{D}$, A is permissible in $\mathcal{D}$ iff the shortfall of A in $\mathcal{D}$ is no greater than the shortfall of any B $\in \mathcal{D}$.[70]

Permissible outcomes in a decision are those which have the lowest shortfall in that decision.[71] The conjunction of [P] and [S] supplies us with a complete moral theory

[67] Note that while Thomas' simple model nevertheless gets around this thorny issue, it does so in a wildly *ad hoc* manner. Plus it does so in a way, as things stand, that violates (†). As will be soon shown, his fuller moral theory does not suffer the same problem.

[68] (Cusbert and Kath, forthcoming, 16, footnote 15)

[69] (Cusbert and Kath, forthcoming, 16)

[70] (Cusbert and Kath, forthcoming, 16); cf. (Kath, 2016)

[71] If so, the permissibility of an outcome is suddenly hostage to what else on finds in his option set. Adding or removing an option can wildly change the state of permissible outcomes in the decision. There's a fork in the road here. Permissibility can be determined by considering either the full option set or by making it dependent on some option set. At first blush, the former seems to be the right way to go. (See especially (Roberts, 2010a, footnote 86, Appendix C).) For one thing, *nontransitivity* in Temkin's sense of the word—see especially (Temkin, 2012, 17); (Cusbert, 2017)—opens the door to getting money-pumped in a sequential choice scenario. For example:

> *Devil's Dutch-Book.* Imagine the Devil offers you the chance to select the next population he brings into existence. For the price of a momentary stay in Hell you can pick between A'B = $\{1, 1, \Omega\}$ and B'C = $\{\Omega, 1, 1\}$. Suppose that you value minimizing comparative harm over spending a moment in Hell. (More so, we are going to assume the population's time in Hell is finite so as to avoid wading into the murky waters of infinite ethics.) After you make your choice, the Devil sneakily introduces another option. He repeats the offer, this time leaving you to choose between keeping your previous choice or switching to AC' = $\{1, \Omega, 1\}$ at the cost of yet another moment spent in Hell. The Devil continues cycling through this series for the rest of your stay in Hell.

So long as the Devil is in charge of what your option set contains, you'll get money-pumped for an eternity of hellfire and brimstone.

To be sure, you might think that this would only be problematic if either toy examples like this had real life counterparts or one couldn't see the money-pump coming. But, in David Lewis' words:

> "the point of any Dutch book argument is not that it would be imprudent to run the risk that some sneaky Dutchman will come and drain your pockets. After all, there aren't so many sneaky Dutchmen around; and, anyway, if ever you see one coming, you can refuse

according to which we ought to *minimize comparative harm to persons* (where comparative harm refers to how much better off they might have been).[72,73]

5.5 Six Candidates

On its lonesome, this type of moral theory does nothing to solve our pesky lil' problem. How does miserable mere addition make the world go worse? Below, I assess the tenability of six different methods for modifying [PS] such that the resulting moral theories are both (a) capable of breaking this silence; and (b) consistent overall with $\mathbb{BN}$.

Of course, there might well be other methods out there that I simply haven't thought of. So I don't dare pretend that the shortlist compiled here is exhaustive. However, the failure of the first five classes of moral theory to satisfy the normative reading of $\mathbb{BN}$ reveals a general theme which I strongly suspect would lead us to reject those other unknown methods too. This realization prompts me to consider a sixth, radically different category of moral theory that does not feature this fatal flaw in §5.5.5; it works.

5.5.1 Better to Have Never Been

Standing at the foot of this dense jungle, the obvious path to try out first is hacking one's way through in a beeline home. Put less colourfully, we only find ourselves in this mess because we earlier subscribed to Existence Noncomparativism. As a first stab, let's abandon this assumption. In a two-outcome comparison, no life at all is better than a miserable life.[74] To this end, let Ω stand for zero.[75]

---

to do business with him. Rather the point is that if you are vulnerable to a Dutch book, whether synchronic or diachronic, that means that you have two contradictory opinions about the expected value of the very same transaction. To hold contradictory opinions may or may not be risky, but it is in any case irrational" (Lewis, 1999, 404-405).

This topic is controversial, and the debate doesn't look to be slowing down. For a recent example, see Ahmed (2017). Money-pumps *qua* counterexamples might indeed be "impotent" to borrow Halstead's term (Halstead, 2015). Perhaps they are. Nevertheless, I will *tentatively* run on the assumption that we should determine the permissibility of an outcome by analyzing it against the whole option set. This option set can be determined by following Roberts' suggestion of considering only what possible worlds are *accessible* to agents (Roberts, 2010a, footnote 86). Later on, when I return to unpack Thomas' moral theory, I drop this assumption.

[72] At (Cusbert and Kath, forthcoming, 16), they rightly point out that this is a satisfactory extension of [B] insofar as (a) any outcome that's impeccable has the absolute minimum shortfall and thereby satisfies [B1]; (b) any outcome that is worse in some respect than an impeccable outcome will have positive shortfall and thereby be impermissible—satisfying [B2].

[73] There may be more than one permissible outcome in a decision.

[74] However, this gets far messier if we also accept a fourth category of value bearers, *undistinguished*, that's up for grabs. See (Gustafsson, forthcoming, 13-14, footnote 18).

[75] A life is often defined as being neutral (in terms of contributive value) if it is neither better nor worse that this life is lived than that it is not lived (Broome, 2004, 142). So, it is reasonable to assign a zero to a 'null life' if that is indeed the neutral lifetime welfare level. But it could be something

It is easy to spot where this goes terribly wrong. Imagine the decision is this: B = $\{\Omega\}$ or B' = $\{4\}$. The Principle of Personal Good tells us that B' is better than B. Therefore, we ought to bring this person into existence. Indeed, it would be impermissible not to, given that B' is the only impeccable outcome. The problems don't end there. Imagine the decision is between: AU' = $\{\Omega, \Omega\}$ or A'U = $\{4, 2\}$. All-things-considered, A'U has the lowest shortfall, and is therefore better than AU'. But this is a combination of a bad thing and a neutral thing. So, intuitively they should have a bad net effect. Of course, both of these implications go against the grain of $\mathbb{BN}$.

So much for our first stab. Let us now consider: How might we alter this moral theory such that we avoid the above implications?

5.5.2 Variabilism

Perhaps the best known example of a moral theory that fits the bill is Melinda Roberts' Variabilism.[76] Her description of the theory is worth quoting in full.

> A middle ground between Inclusion and Exclusion is Variabilism. The fact that Variabilism is at least plausible means that, having rejected the frying pan, we should not feel immediately compelled to leap into the fire.
>
> Variabilism can be put as follows.
>
> *Variabilism:*
>
> A loss incurred at any world w by any person p has moral significance for purposes of determining the permissibility of any act a performed at w that imposes that loss and any alternative act a' performed at any accessible world w' that avoids that loss *if and only if* p does or will exist at w.
>
> By implication, the loss incurred by p at w when agents fail to bring p into an existence worth having at w will have no moral significance for purposes of determining the permissibility of the act a performed at w that imposes that loss or any alternative act a' performed at w' that avoids that loss.[77,78]

In a nutshell, if someone p does not exist at world w, then the loss he suffers can be ignored for it has no moral bite.

But there are going to be variable population decisions where a morally relevant loss will obtain no matter what we do. In these kinds of cases, we require some well-defined procedure which describes how these tradeoffs are supposed to be handled.

---

other than zero if we adopt, for example, a critical-level threshold. (Or it may not have any constant numerical value—see (Broome, 2004, 143).)

[76] Views of this sort can be traced back to (McDermott, 1982). Theories in the same ballpark include (Meacham, 2012) and (Ross, 2015).

[77] (Roberts, 2010a, 76)

[78] By Inclusion and Exclusion Roberts is referring us to opposing positions on the moral issue of whether merely possible persons ought to factor into our evaluations. As the names suggest themselves, the former includes them, the latter excludes them.

Though there might be some number of ways to unpack them even further, there are roughly two such procedures available to us.

On the first, an outcome's shortfall is equal to the total lifetime welfare of the persons in w minus the total lifetime welfare of the same persons in w', where w' is an available world in which the w-persons have greater total welfare than they do in any other available world. The outcome with the lowest shortfall-score is permissible, and outcomes with a higher shortfall-score are impermissible. Call this the Total-Sum-of-Losses procedure (hereafter abbreviated '$\mathbb{TSL}$').[79]

On the second, an outcome's shortfall is equal to the sum of every individual's lifetime welfare in w minus the welfare he would have had if the best outcome for him obtains. The outcome with the lowest shortfall-score is permissible, and outcomes with a higher shortfall-score are impermissible. Call this the Sum-of-Individual-Losses procedure (hereafter abbreviated '$\mathbb{SIL}$').

It can now be shown that Variabilism is incompatible with $\mathbb{BN}$ under either $\mathbb{TSL}$ or $\mathbb{SIL}$.

5.5.2.1 Objecting to $\mathbb{TSL}$

Consider:

> *Fortune-Teller's Admonition.* Imagine a fortune-teller presents you with a button which whenever pressed brings a happy person into existence somewhere in a disconnected part of the galaxy. This person would be a mere addition. You quickly surmise that you are permitted to perform either action in accordance with $\mathbb{TSL}$. The fortune-teller, however, has peered into his crystal ball in the meanwhile. Sullenly, the fortune-teller declares that he sees an infant in one of your future timelines which suffers from a horrible disease. The infant's life will be, on balance, just barely not worth living. Yet, this little bundle of misery will enrich your life, even going so far as to make you a better person. Should you now press the button?

Outcome	$\mathbb{TSL}$-Score
L = (10, Ω, Ω)	10 11 = 1
LM = (10, 5, Ω)	15 15 = 0
L'N = (11, Ω, 2)	9 10 = 1

Table 5.1: Fortune-Teller's Admonition

The only permissible action left for you to perform, once the fortune-teller has peered into his crystal ball, is mere addition. You ought to bring about LM.

The general lesson here is that, for any conceivable mere addition case, I can approach the deciding agent and *make it the case* they ought to press the button. I do not

[79] This is based on Daniel Cohen's original formulation of Expected Actualism. See (Cohen, forthcoming).

even require the power of a crystal ball; it is sufficient that I am telling the truth when I claim there is such a possible world—at first blush, this seems like a very low bar to meet. But you should do this, on our moral theory, not because this mere addition's life would be worth living, but for the very strange reason that I merely reported to you the possiblity of *some* possible world in which you go on to perform benign miserable addition. Obviously, $\mathbb{TSL}$ is untenable.

5.5.2.2 Objecting to $\mathbb{SIL}$

Consider:[80]

Outcome	$\mathbb{SIL}$-Score
E' = (1, Ω)	1 10 = 9
E"I = (10, 10)	0 + (90) = 90
EI' = (200, 100)	210 + 0 = 210

Table 5.2: Bonkers Conclusion

The only permissible outcome is E'. So, this is the outcome you ought to bring about. This however means that we are prohibited from performing benign addition. This too goes against the grain of $\mathbb{BN}$.

5.5.3 Schwartz's Method

The general idea of harm-minimization, in the sense advocated by Roberts, McDermott, and others, seems to be tracking on to the right 'moral stuff'. But it cannot satisfy $\mathbb{BN}$ as it stands either way we cash things out. Whereas $\mathbb{TSL}$ required mere addition (and $\mathbb{SIL}$ does not), $\mathbb{SIL}$ tells us that benign addition is impermissible (and $\mathbb{TSL}$ does not). Myself, I would not be willing to bite either of these bullets.

So let's rethink our approach. Specifically, it can be altered so that the shortfall of an outcome is determined in piecemeal fashion, with only two outcomes being compared at a time rather than across the full option set. This means that the outcome which does best in some morally relevant respect is confined to one of the two outcomes under comparison at any given time. Afterwards, Thomas Schwartz's method will be applied to determine the overall (im)permissibility of every outcome in the set.

The basic idea is simple enough. As before, we want to minimize shortfall. We will require a new piece of terminology. For a given decision, $\mathcal{D} \subseteq \mathbb{A}$, define a non-empty subset $A \in \mathcal{D}$ as being *deliberatively stable* if no available option $B \in \mathcal{D}$ has lower shortfall in a two-way comparison with A. According to Schwartz's method, an option is permissible just in case it is contained in a *minimal* deliberatively stable subset of

[80] Jacob Ross presented this toy example at the Conference on Theoretical Population Ethics, Oxford University, 21-22 November 2015. I'm unaware of any manuscript or paper in which he has since published the idea (cf. (Visak, forthcoming)).

Figure 5.3: The diagrams illustrate the results according to $\mathbb{TSL}$. On the left you will find *Fortunes Revisited. Bonkers Once More* is on the right. The results are essentially the same for $\mathbb{SIL}$, with the only small difference being that in *Fortunes Revisited* the harpoon from L to L'N is headless. When reading these diagrams, a harpoon pointing from an outcome A to another B represents A as having a lower shortfall relative to {A, B}. The headless harpoon represents equality.

the available options.[81,82] If the minimal deliberatively stable subset is empty, as is the case if the overall ranking is intransitive, he maintains that realizing any of the available options is permissible.

Armed with Schwartz's method, we can now re-analyze both $\mathbb{TSL}$ and $\mathbb{SIL}$.

Comparison	$\mathbb{TSL}$-Score	$\mathbb{SIL}$-Score
{L, LM}	L = 0 LM = 0	L = 0 LM = 0
{L, L'N}	L = 1 L'N = 1	L = 1 L'N = 0 + (2) = 2
{LM, L'N}	LM = 0 L'N = 1	LM = 1 + 0 = 1 L'N = 0 + (2) = 2

Table 5.3: Fortunes Revisited

Comparison	$\mathbb{TSL}$-Score	$\mathbb{SIL}$-Score
{E', E"I}	E' = 9 E"I = 0	E' = 9 E"I = 0
{E', EI'}	E' = 0 EI' = 120	E' = 0 EI' = 201 + 0 = 201
{E"I, EI'}	E"I = 0 EI' = 120	E"I = 0 + 90 = 90 EI' = 210 + 0 = 210

Table 5.4: Bonkers Once More

The minimal deliberatively stable subset in Fortune-Teller's Admonition is {L, LM}, and it is {E', E"I} in Ross' Bonkers Case. So while we are no longer gored on the fortune-teller's horn, both versions of Variabilism get the Bonkers Case wrong. E"I is better than E' by [R3]. There is yet more stormy weather on the horizon for both.

[81] (Thomas, 2016, 210)
[82] See (Schwartz, 1972), as well as (Ross, 2015, §5).

Figure 5.4: Diagram for Devilish Proposal. $\mathbb{TSL}$ is on the left, and $\mathbb{SIL}$ on the right.

Consider:

> *A Devilish Proposal.* Imagine the devil is offering to let you make his decision concerning which population he will bring about in Hell. In a rare slip on his evil part, one of the available outcomes belongs to $\mathcal{H}$.

M' = (2, Ω),
M" = (1, Ω),
MH' = (3, 10000),
H = (Ω, 9994).

$\mathbb{TSL}$ gets this one wrong (at least so far as $\mathbb{BN}$ goes), and things pan out even worse for $\mathbb{SIL}$.

Comparison	$\mathbb{TSL}$-Score	$\mathbb{SIL}$-Score
{M', M"}	M' = 1 M" = 0	M' = 1 M" = 0
{M', MH'}	M' = 0 MH' = 0	M' = 0 MH' = 1 + 0 = 1
{M', H}	M' = 2 H = 0	M' = 2 H = 0
{M", MH'}	M" = 0 MH' = 0	M" = 0 MH' = 2
{M", H}	M" = 1 H = 0	M" = 1 H = 0
{MH', H}	MH' = 0 H = 3	MH' = 3 H = 6

Table 5.5: Devilish Proposal

Using Schwartz's method on $\mathbb{TSL}$ results in only MH' being permissible. $\mathbb{SIL}$, on the other hand, results in an intransitive ordering; thereby, all four options are equally permissible. This is insane. We should take advantage of the Devil's rare slip, and bring the only happy outcome, H, about. After all, every other outcome involves a bad thing or the combination of a bad thing and a neutral thing—so, $\mathbb{BN}$ tells us that they must each be overall bad.

5.5.4 Regret Minimization

As before, it would seem as if we are onto something following this train of thought. So let's dig a little deeper.

Suppose that we are prepared to loosen up the unshakable stance held by $\mathbb{BN}$. Let's instead maintain that a bad thing combined with a neutral thing can, *under the right conditions*, be a neutral thing all-things-considered. In a nutshell, the badness of miserable addition can be neutralized if their suffering is outweighed by the lifetime welfare which would obtain for the other, non-miserable additions. Following this recipe, the addition of MH' to $\mathrm{B} \in \mathcal{H}$ would be of neutral contributive value.[83]

But, this being said, there is one wrinkle so as to preserve the bite of $\mathbb{BN}$. MH' cannot be better than some other outcome belonging to $\mathcal{H}$. Both MH' and H ought to be permissible in the Devilish Proposal according to this revamped interpretation of $\mathbb{BN}$. The reasoning unpacked is this: we might be willing to look the other way about such a state of affairs obtaining (after all, on balance it is not such a horrible thing); yet, we should never be obligated to bring it about.

We already have the basic building blocks of such a moral theory set in place. $\mathbb{TSL}$ and $\mathbb{SIL}$ can be supplemented with $\mathbb{CN}$. Thomas refers to the former as *Regret Minimization*.[84] He does not consider the latter, but we can call it *Personal Regret Minimization* after him.

Regret Minimization has the following structure.[85] For every two outcomes being compared A and AB, there is a value of AB for the persons in A, denoted $V_{\mathrm{A}}(\mathrm{AB})$. Given $\mathcal{D}$, the regret of A, Reg(A), is the extent to which the total sum of welfare in A falls short of the option AB that does best in terms of the sum of *their* welfare:

$$\mathrm{Reg}(\mathrm{A}) = \min_{\mathrm{AB} \in \mathcal{D}}((V_{\mathrm{A}}(\mathrm{AB}) \;\; V_{\mathrm{A}}(\mathrm{A})).$$

Outcomes that minimize regret are permissible on this model.

$V_{\mathrm{A}}(\mathrm{AB})$ can now be defined in terms of its theoretical backbone, $\mathbb{CN}$. If there are fewer contingent persons in AB than in A, then $V_{\mathrm{A}}(\mathrm{AB})$ is just the total welfare in AB. If there are more contingent persons in AB, though, then $V_{\mathrm{A}}(\mathrm{B})$ is the necessary persons total welfare, plus the total welfare that the contingent persons in A would have if they were treated as having the same average welfare as the contingent persons in B. In symbols:

$$V_{\mathrm{A}}(\mathrm{AB}) = \begin{cases} \mathcal{W}(\mathrm{A}) + \mathcal{W}(\mathrm{B}), & \text{if } \mathcal{N}(B) \leq \mathcal{N}(A) \\ \mathcal{W}(\mathrm{A}) + X, & \text{otherwise.} \end{cases}$$

Bear in mind, if the contingent persons belonging to B have lives not worth living, then this fully counts against outcome AB. For simplicity, we will assume that individuals

[83] You will recall, this amounts to adopting what I previously called the weak version of the Non-Greedy Principle of Personal Good.

[84] (Thomas, 2016, 242ff)

[85] Throughout my explanation I am quoting material from (Thomas, 2016, 244-245).

composing B either all have lives worth living or not. They're in the same boat so to speak. This gets us:

$$X = \begin{cases} \mathcal{W}(\mathrm{B}), & \text{if } \mathcal{W}(\mathrm{B}) \leq 0 \\ \mathcal{W}(\mathrm{B}) \cdot \frac{\mathcal{N}(A)}{\mathcal{N}(B)}, & \text{otherwise.} \end{cases}$$

In a decision {AB, AC}, Personal Regret Minimization determines Reg(AB) by the extent to which the total sum of *individual regrets* in AB falls short of the option AC that does best in terms of the sum of *their* regrets. In those cases where $\mathcal{N}(\mathrm{C})$ is greater than $\mathcal{N}(\mathrm{B})$, the counterparts of $b \in B$ in outcome C are treated as if their welfare were the same as the average welfare of the contingent persons in C. The regret of every $b \in B$ is then independently scored against their own best, whether this be the welfare level they have in outcome AB or the average lifetime welfare level of $c \in C$ that's derived in AC. Let ϕ represent this individual 'regret-score'.

Expressed in symbols:

$$V_{\mathrm{AB}}(\mathrm{AC}) = \begin{cases} \left(\sum \phi_{a \in A}\right) + \left(\sum \phi_{c \in C}\right), & \text{if } \mathcal{N}(C) \leq \mathcal{N}(B) \\ \left(\sum \phi_{a \in A}\right) + X, & \text{otherwise.} \end{cases}$$

X can be now defined as follows:

$$X = \begin{cases} \sum \phi_{c \in C}, & \text{if persons belonging to C have lives not worth living} \\ \sum^{\mathcal{N}(B)} \phi_{c \in C}, & \text{otherwise.} \end{cases}$$

Both Regret Minimization and Personal Regret Minimization fail to satisfy our revamped $\mathbb{BN}$ in the Devilish Proposal so long as Schwartz's method is still being applied. Indeed, we find the exact same results as before (see next page), with $\mathbb{TSL}$ finding only MH' permissible, and $\mathbb{SIL}$ allowing for anything goes. But this problematic feature can be, of course, removed. Re-running the analysis with our initial method—a direct-rule for decisions involving more than two options—we find that under $\mathbb{SIL}$ the right results come out (see next page).

Alas, they only come out right on this one toy example. If we were to make the tiniest alteration to MH' such that it were $(2, 10000)$, then H becomes impermissible once more. Plus, $\mathbb{SIL}$ so-understood generates the claim that only E' is permissible in the Bonkers Conclusion.[86] It furthermore determines that only L is permissible in a Fortune-Teller's Admonition.[87] Given the progress we have made so far in the paper, one might (understandably) be inclined to just weaken our revamped $\mathbb{BN}$ even further, in fact all the way down.

Perhaps MH' should be the only permissible option in the Devilish Proposal. While

[86] It ranks their shortfall scores, respectively, as follows: $9; 90; 210$.
[87] It ranks the shortfall scores, respectively, as follows: $0; 1; 7$.

Comparison	$\mathbb{TSL}$-Score	$\mathbb{SIL}$-Score
{M', M"}	M' = 1 M" = 0	M' = 1 M" = 0
{M', MH'}	M' = 0 MH' = 0	M' = 0 MH' = 1 + 0 = 1
{M', H}	M' = 9996 H = 0	M' = 9996 H = 0
{M", MH'}	M" = 0 MH' = 0	M" = 0 MH' = 2 + 0 = 2
{M", H}	M" = 9995 H = 0	M" = 9995 H = 0
{MH', H}	MH' = 0 H = 6 + 3 = 3	MH' = 3 + 0 = 3 H = 6 + 3 = 3

Table 5.6: Regret-Min on Devilish Proposal

Outcome	$\mathbb{TSL}$-Score	$\mathbb{SIL}$-Score
M' = (2, Ω)	2 9994 = 9996	2 9994 = 9996
M" = (1, Ω)	1 9994 = 9995	1 9994 = 9995
MH' = (3, 10000)	9997 9997 = 0	(10000 - 10000) 3 = 3
H = (Ω, 9994)	9994 9997 = 3	9994 (10000 3) = 3

Table 5.7: The Devil Refused

H may well be the only happy outcome available, it is not as if $h \in H$ does not suffer a morally relevant loss. If that's where one finds himself at this late hour, then they are free to put the matter to bed for now—$\mathbb{TSL}$ under either Schwartz's method or without it but supplemented with $\mathbb{CN}$ does the trick.

I am however equal parts restless and stubborn, and not yet ready to give up. A little more tinkering will, I'm sure, culminate in a coherent moral theory that satisfies the strong normative reading of $\mathbb{BN}$; so, my search continues.

5.5.5 Grrr! One More Try... A Multi-Step Framework

There may be countless more variations on this basic model that could be tested. However, we might save ourselves some grief by instead picking up on a recurring theme that arose while toying around with Cusbert and Kath's proposed multidimensional model.

$\mathbb{BN}$ requires several morally relevant respects to function in unison, but they are opposing forces that often push up against each other. The badness of adding a miserable person to the population must be set against the goodness of lifting an existing member's lifetime welfare up, as well as against the goodness of bringing about the state of affairs that is better for persons being added to the population, and so on.

Attempting to fit all of these items together in one fell swoop has over and over ended up with random bits poking out. So here is my thought. Don't have them fight over the scraps all at once like a pack of hyenas. Rather, let them loose a few at a time on the 'carcass' in the order that matters most (according to $\mathbb{BN}$).

Of course, this means that there is potential here for the last set loose to go hungry. But I think this is a small price to pay if we are committed to getting $\mathbb{BN}$ going. Plus, we might as well see what falls out of this proposal, given we have already come so far.

More exactly, here is what I propose.

While holding onto [P], let's substitute the following bridge principle in for [S]:

> [M]. For all decisions $\mathcal{D}$ and all A, B $\in \mathcal{D}$:
>
> > [M1] If the shortfall$_{nec}$ of A $\in \mathcal{D}$ is no greater than the shortfall$_{nec}$ of any B $\in \mathcal{D}$, then A is *initially permissible* in $\mathcal{D}$.
> >
> > [M2] From the set of initially permissible outcomes in $\mathcal{D}$ derived by [M1], A is *decidedly permissible* in $\mathcal{D}$ iff the shortfall$_{con}$ of A is no greater than the shortfall$_{con}$ of any initially permissible option B $\in \mathcal{D}$.

Here shortfall$_{nec}$ stands for the total sum of shortfalls suffered by necessary persons, where a person counts as being necessary only if they exist in every outcome in a given decision. By contrast, shortfall$_{con}$ is the total sum of shortfalls suffered by contingent persons.

Bear in mind, we adhere by Roberts' ruling which holds that losses only morally matter if they are suffered by an existing person. If the person does not exist in an

outcome, then whatever positive shortfall he might suffer does not factor into the equation. Moreover, following Thomas' suggestion, even if a person in B were contingent, so long as his life is not worth living, this should count against the initial permissibility of B. In other words, at the first step we sum up the shortfall of both necessary persons and miserable contingent persons.[88] Those outcomes with the highest score at the end of step one are considered *initially permissible*. Outcomes that are not initially permissible are *decidedly impermissible*. If there is only one initially permissible outcome, then it becomes *decidedly permissible*.

The remaining subset of persons in B are accounted for at step two of our procedure [M] *only if* there is more than one initially permissible outcome. Here we subtract the individual shortfall of those persons that we have yet to account for—where, importantly, the outcome that does best with respect to p is limited to those outcomes which are initially permissible in $\mathcal{D}$—from the score of the intitially permissible outcomes. Step two acts as a tie-breaker so to speak. If several outcomes remain which share the highest score after executing step two of [M], then the remaining options are all decidedly permissible. All other initially permissible outcomes with lower final scores are decidedly impermissible.

The proposed moral theory is inelegant, no doubt. But this need not imply that it's, as such, ad hoc. It's no longer obvious to me that, in terms of satisfying the normative reading of $\mathbb{BN}$, axiological neutrality alone could be sufficient grist for that windmill. This is what the paper's analysis suggests, at any rate. So we must intervene on a different bit of moral machinery to accurately capture the spirit of $\mathbb{BN}$. A good or bad thing combined with a neutral thing should not have the net effect of being morally neutral. Bearing this in mind, the above system can be described as giving the moral upper-hand at the level of permissibility to good and bad things over neutral things. There is nothing contrived about the moral theory so long as we understand this as the primary stimulus behind it.[89]

The tradeoff-tyranny over neutral things, though, cannot totally settle the matter in all cases on our moral theory. If there are some number of outcomes that are equally permissible with reference to the good and bad things, then I think these neutral things should be afforded the ability to evenutally tip the scale, such that they determine whether initially permissible outcomes are also decidedly permissible. Even if we are neutral about additional lives worth living when assessing which subset of $\mathcal{D}$ is initially

[88] Although I have not opted to do so, one may, if he wishes, break this up into two steps. At the first we take the sum of shortfalls incurred by non-uniquely realizable persons in this outcome. At step two we add the lifetime welfare of lives not worth living from this initial score for all of the remaining initially permissible outcomes. Step three then resolves any ties by allowing the shortfall of extra lives worth living to settle the matter. Alternatively, steps one and two coud be reversed. This would be the case if one believed that there were absolutely no tradeoffs that could morally justify either creating a life not worth living or reducing a person to utter misery. (cf. (Brown, 2005) and (Brown, 2007))

[89] Alas, the moral system so-described supplies no ammunition against Jeff McMahan's challenge. See (McMahan, 1981); (McMahan, 2009, 54). In this sense, it is still ad hoc. But this is not ad hocery inherent to my solution; rather, this ad hocery is built right into the supporting structure of $\mathbb{BN}$: its core assumptions about the modal operator as being relevant to defining the role played by goods, as well as evaluating the contributive value of these goods when contained in a life lived.

permissible, we are not moral monsters. If some of the outcomes we have hived off in step one would be worse for extra persons, then this fact should militate against their being overall permissible.

This brings order and civility to what is otherwise, as I colourfully described earlier, a pack of hyenas viciously pawing and snapping at each other over a carcass.

5.6 Analysis

The moral theory that we are left with has a well-behaved axiological basis. It is internally consistent, and adequately describes the betterness relations between outcomes as required by the evaluative reading of $\mathbb{BN}$.

Of course, we took a few small deviations from the axiological framework as initially described. For starters, we abandoned the overall betterness framework for a multidimensional framework. An outcome, on the adopted model, can be better or worse than another outcome in several morally relevant ways. This means that our rankings might be incomplete, given that some outcomes could be incomparable with each other. We next found ourselves forced to claim that non-existence is comparable with a good, neutral, or bad life in terms of being worse for, same as, or better for this person. Although I have chosen not to go ahead and incorporate $\mathbb{CN}$, someone could do so if they wished. This would then be yet another cost to bear in mind. These are the obvious axiology-related divergences from Broome's original project.

There is a less obvious cost which does not count against the moral theory, though, insofar as the final package still bears similarity to our original adumbration of $\mathbb{BN}$'s normative commitments. Nonetheless, it is a cost worth highlighting. Our assessment of the varying shortfalls in a decision is cut off at the knees in two senses on my model. First, we ignore comparative harms to happy contingent persons at the first step of the procedure. Second, a contingent person's shortfall is defined with reference to the best outcome for him which is still standing at step two. Consider the following decision:

$$\begin{gathered}\text{LM'} = (2,\ 8),\\ \text{L'} = (4,\ \Omega),\\ \text{L'M} = (4,\ 3).\end{gathered}$$

At step one LM' is found to be decidedly impermissible. But compare LM' with L'M. LM' is plausibly better than L'M—after all, it has higher total lifetime welfare (and a lower total sum of shortfalls). If it weren't for L' being part of the decision, LM' would have been decidedly permissible according to our moral theory, and L'M would have been decidedly impermissible. For some of us, this deviation from the program will stick in their craws. I confess that I too find it a bit strange, morally speaking. However, this also strikes me as an unavoidable move towards satisfying $\mathbb{BN}$.

This being said, a cursory review might prompt one to claim that the above feature, in fact, *does* involve an unacceptable axiological cost: giving up on impartiality. This is not an unreasonable objection, given there is most definitely some inkling of partiality at work here. But this is a mistake. Consider:

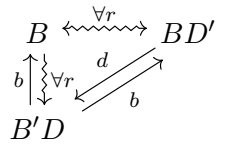


Figure 5.5: Still Impartial Between BD' and B'D

$$\begin{gathered} \text{B} = (2, \Omega), \\ \text{BD'} = (2, 4), \\ \text{B'D} = (4, 2). \end{gathered}$$

BD' is better for d, and B'D is better for b. Neither is overall better or worse than the other. Nor are these outcomes overall equally as good as each other. See Figure 5. Strictly speaking, this is a violation of Broome's principle, Impartiality Between People.

But this is hardly out of left field. Impartiality Between People was formulated with an overall betterness framework in mind. It doesn't accurately track for what is going on in a multidimensional framework. So, we must reconsider what counts as preserving impartiality in such a context. It seems to me that we are being appropriately impartial by acknowledging that (BD' $\succ_d^2$ B'D) and (B'D $\succ_b^2$ BD'). Sure, the two outcomes are not overall equally as good. Yet, one is just as bad for b as the other is bad for d; and we are not committed to describing BD' as worse in terms of axiological value just because b exists in every outcome captured by the wider option-set {B, BD', B'D}. What we are committed to, on the proposed moral theory, is the following claim: *insofar as b is a member of the original population, permissiblity is determined by giving him the upper-hand over d.* In other words, we are only being partial towards b at the normative level. This is consistent with both impartiality at the evaluative level and the normative reading of $\mathbb{BN}$.

It can now be established that [M] is a consistent extension of [B]. Decidedly permissible outcomes according to [M] are those that have the lowest shortfall in the subset of outcomes that are initially permissible. [B1] is satisfied for the reason that an impeccable outcome will be both initially and decidedly permissible. It will survive through both steps of our procedure because there is no respect in which it is worse than any other outcome B $\in \mathcal{D}$. [B2] is also satisfied, since an available outcome that is worse in some respect than an impeccable outcome—*either* for the original population as well as miserable persons *or* at step two for extra persons—thereby has positive shortfall, and will therefore be proclaimed decidedly impermissible.[90]

*
* *

[90]Here I have offered a quick restatement of Cusbert and Kath's proof that [S] extends [B] in an adequate manner, except I have substituted [M] for [S]. The proof is essentially the same for both extensions of [B].

To close, I will now demonstrate that [PM] describes those outcomes in the seven toy examples covered in the chapter in a way that is consistent with the normative reading of $\mathbb{BN}$.

Outcome	*Step 1*	*Step 2*	*Ruling*
A' = (6, Ω)	0	0	permissible
A'B = (6, 1)	0	(0 0) = 0	permissible
AB' = (4, 4)	2	✗	✗
A = (4, Ω)	2	✗	✗

Table 5.8: Broome's Toy Example

In Broome's Toy Example, mere addition is not required but permissible. The remaining outcomes are impermissible because they involve either violation of the Principle of Personal Good or the Non-Greedy Principle of Personal Good.

Outcome	*Step 1*	*Step 2*	*Ruling*
A = (1, Ω)	skip	0	permissible
B = (Ω, 1)	skip	1	✗
B' = (Ω, 2)	skip	0	permissible

Table 5.9: Nonidentity Problem

We skip the first step in the Nonidentity Problem case, since there are no non-uniquely realizable persons to factor in in this case. In accordance with the Principle of Personal Good, B is found to be impermissible. A and B' are both permissible.[91]

Outcome	*Step 1*	*Step 2*	*Ruling*
AB' = (2, 5, Ω)	(2 + 0) = 2	✗	✗
ABC' = (2, 4, 4)	(2 + 1) = 3	✗	✗
A'BC = (4, 4, 2)	(0 + 1) = 1	(1 0) = 1	permissible

Table 5.10: (Im)Partiliaty Case

As demonstrated in Table 5.10, we find that ABC' is impermissible in the (Im)Partiality Case given that it required making b worse off. We also find that AB' is impermissible. This is because a's shortfall is two in AB', while b's shortfall is 1 in A'BC. Although c is two units worse off in A'BC than he would be in ABC', his individual shortfall is calculated by turning to the outcomes remaining that are best in this respect. Since there is only one remaining outcome after performing step one, his shortfall is 0, not 2.

In the Devil's Dutch-Book every outcome has an initial score of 1. This is because every outcome involves a miserable person. The shortfall of each person at the second step is also the same in every outcome. Therefore, all three outcomes are permissible.

[91] If we adopted $\mathbb{CN}$, only B' would be decidedly permissible. Again, I am less bothered by this result insofar as my goal is strictly to salvage the (stronger) normative reading of the $\mathbb{BN}$.

Outcome	*Step 1*	*Step 2*	*Ruling*
A'B = (1, 1, Ω)	1	(1 0) = 1	permissible
B'C = (Ω, 1, 1)	1	(1 0) = 1	permissible
AC' = (1, Ω, 1)	1	(1 0) = 1	permissible

Table 5.11: Devil's Dutch-Book

Outcome	*Step 1*	*Step 2*	*Ruling*
A = (10, Ω, Ω)	1	1	permissible
AB = (10, 5, Ω)	1	(1 0) = 1	permissible
A'C = (11, Ω, 2)	(0 + 2) = 2	✗	✗

Table 5.12: Fortune-Teller

Both A and AB in Fortune-Teller involve making a non-uniquely realizable person worse off. A'C, though, involves bringing about a life not worth living. The shortfall attributed to this miserable life is slightly worse than the shortfall of a in either A or AB. Therefore, since b's loss in A has no moral bite, and this loss is 0 in AB, both outcomes are permissible.

Outcome	*Step 1*	*Step 2*	*Ruling*
A' = (1, Ω)	9	✗	✗
A"B = (10, 10)	0	(0 0) = 0	permissible
AB' = (200, 100)	210	✗	✗

Table 5.13: Bonkers Case

At step one we find both A' and AB' to be decidedly impermissible in the Bonkers Case. The latter involves (extreme) malign addition. The former is worse because a is better off in A"B. Even though b is 90 units worse off in A"B than he would be in AB', the shortfall is determined to be 0 on our model.

Finally, in the Devilish Proposal we find that although there are no non-uniquely realizable persons in this decision, all but B involve miserable persons. Performing step one of [M] governs that only B is initially permissible. This is consistent with our hard line position on the prohibition on miserable mere addition.

It would seem that at last we have arrived at a well-behaved axiology and bridge principle that together produce a moral theory that achieves our goal of satisfying the normative reading of $\mathbb{BN}$. Perhaps the basic idea of this paper will have opened the door to further proposals down the line. The war, far be it from lost, rages on.

5.7 Conclusion

It beehoves me to now explain what this all means in terms of the go-fast policy on my toy model.

Outcome	*Step 1*	*Step 2*	*Ruling*
A' = (2, Ω)	2	✗	✗
A" = (1, Ω)	1	✗	✗
AB' = (3, 10000)	3	✗	✗
B = (Ω, 9994)	skip	0	permissible

Table 5.14: Devilish Proposal

Going fast means that the generation currently alive should develop in a way that maximally increases their own basket of 𝔼s unless doing so would bring about a broken world of sufficient severity. I am going to make three assumptions in my toy model to capture the minimal safety net that go-fast requires being put in place, given both growing cumulative man-made catastrophic risk and the possibility of a freak catastrophe rocking their world at the drop of a hat. One assumption is simply made so as to avoid bedevilling the toy model. The remaining two assumptions also happen to serve that purpose, but are very, very charitable to the go-fast policy, and my case is, as such, more difficult to make.

Let's start with the charitable assumptions. (a) I'll assume that avoidable (man-made) catastrophic risk can be reliably forecasted and measured. Before it runs too high and doomsday is triggered, the go-fast policy requires a final happy generation be brought into existence—bear in mind, the population guarantees that this is the final generation by exhausting their remaining energy pool by creating lots of happy persons. (b) I assume that if a freak catastrophe hits and mankind survives, then all those currently alive and every generation subsequently brought into existence is permanently reduced to 𝔹, a life just barely worth living. There would be, for example, no chance of recovering from a freak gamma-ray burst, no matter how much time were to pass.

Now, one might ask himself: "what about step two of [M]?" He doubts that (b) is as charitable as I am making it out to be. He would continue his thought: "should step two of [M] not kick in, such that if subsequent generations could have been better off under a better safety net, then go-fast would have required it?" *In theory:* yes; so long as the current generation is equally well off in both outcomes, then step two dictates that the outcome in which subsequent persons' shortfall is higher is impermissible. *My sterner reply:* but the line in the sand has to be drawn somewhere! On my toy model, the go-fast development speed is the same as in the best case scenario described at the outset of the chapter. It would be giving go-fast its cake and letting them eat it too for me to *also* allow them to have a perfectly good safety net (such that they recovered just fine) which their forebears paid nothing for (where the cost is developing slower). And if they did have to pay something for it, then this outcome is worse than the one in which they did not—this is to say, the happier outcome for future persons would have been ruled out at step one of [M].

Someone could continue to press me: "why keep them stuck at 𝔹 instead of some random value of 𝔼s which is low, but positive?" In other words, why assume that every

freak catastrophe, no matter its scope or severity, results in the same bleak state of affairs. My reasoning for this is essentially the same as above. Already I have committed us to the wild assumption that adherents of go-fast are omniscient when it comes to man-made catastrophes. Next, I charitably assumed that the lives of survivors would be worth living *despite the fact their forebears did nothing at all to that end*.[92] But I did not have to make this assumption. After all, (*i*) the benefits of breaking the world might be so great for the current generation as to swamp the bad thing of breaking the world for some number of future persons. If that were so, then go-fast would say it is a good thing all-things-considered. I have ignored the possibility of such tradeoffs even though it would make my own case for go-slow that much easier. (*ii*) Axiologically, go-fast ranks extinction over the broken world outcome. But in practice the go-fast policy would, under conditions of uncertainty, allow the present generation to roll the dice if the benefits for themselves were sufficiently large. So, really, both the number of terrible lives that go-fast might allow subsequent persons to suffer and the extent of their wretchedness is considerably worse than the fate of a similar tradeoff permitted under conditions of certainty. Take (*i*) and (*ii*) together, as I really should have just done in the first place, and it's crystal clear that it is go-fast that took me to the cleaners, not the other way around!

Bearing the above in mind, I stand by my second assumption, and, indeed, confidently report that it is overall a conservative assumption. Again, (b) while survivors of a freak catastrophe will not suffer a fate worse than death, they will have neither the capabilities nor necessary resources with which to climb out of the $\mathbb{B}$-Hell their forebears placed them in by following go-fast. My third assumption has already been stated, as it's built into my reasoning for (b), and it is this: (c) the minimal safety net provided by go-fast costs their population nothing.

[92] In fact, I don't make the same assumption with go-slow, and they plainly *deserve it* more than go-fast!

6

Results: *Safety-First Wins The Race*

> A bear, however hard he tries, grows tubby without exercise.
>
> A. A. Milne, *Winnie-the-Pooh*

This short chapter presents the results of my toy model. You will recall, there are four general categories which exhaust the range of possible outcomes on my toy model:

(a) *go fast and suffer a broken world*;

(b) *go fast and early extinction*;

(c) *go safe-n-slow and early extinction; or*

(d) *go safe-n-slow and prolong humanity's place among the stars (with the occasional freak catastrophe temporarily lowering their lifetime welfare).*

I argued in chapter 4 that (a) is worse than (b), and (c) is worse than (d) according to the Competing Claims View. But because there are too many random influences on the population's evolutionary trajectory, there are countless outcomes which vary continuously in their goodness on both safe-n-slow and the dangerously fast policies. So, we could not determine whether (b) was better or worse than (d) all things considered. We resolve this difficulty in this chapter by running the toy model on `Python`. I'll demonstrate that a variant of safe-n-slow, the *safety-first* lottery, is what does best in expectation.

I begin by recapping the key lessons of previous chapters. The precise design of the toy model is explained in section 6.2. There I'll furthermore clarify how my design is adapted to each policy choice. Section 6.3 goes on to summarize the results of running the model. Again, these results describe how things tend to get coloured in, given some policy choice. The collection of these simulations is a robust sample of the larger

sample space of outcomes which we could not adequately address from our armchairs. The chapter will close by offering an assessment of the results. What we find is that safety-first is the correct development policy for the real world.

6.1 Recap

Let's start by reviewing what the Competing Claims View requires in terms of population policy. In chapter 4 I argued that under either development policy the size of a generation ought to be fixed at some lower-bound. Any smaller and this generation might not survive a disaster—e.g., a bad case of the flu could snuff out humanity if it's just Adam and Eve.[1] Plus their development will be slower. There is some balance of adding just enough persons so that their promotion of $\mathbb{E}$s is more efficient without jeopardizing humanity's place among the stars too much. I have not tried to determine this balance in the dissertation.

I also explained why overpopulation should be avoided on grounds of fairness in chapter 4. The fact that overpopulation is bad for present persons—e.g., food shortages—isn't something I have tried to account for in the model. Rather, I stipulated that no such harms would befall the guilty generation, and then argued that overpopulation is nevertheless wrong because *it places some subsequent generation(s) in the trilemma of choosing between (a) having to lower their own size in order to maintain an equal distribution going forward; (b) exhausting the remaining pool of energy by maximally propagating; or (c) leaving the next generation with an even harsher cut to make (in terms of (a))*. To be sure, this does not mean that a generation cannot create enough persons to collectively devour the remaining energy pool. Because this act would prevent subsequent generations from being brought into existence, this generation does not qualify as overpopulated on my model. This, I went on to claim, is largely consistent with the demands of fairness insofar as we aren't accountable to persons that will never exist—they just aren't capable of bearing interests or making claims on us. However, I did go on to argue that there is nevertheless something unfair about this outcome under certain conditions. So to speak, if previous generations have endured great sacrifices in effort of salvaging humanity's future, then we do them *some* injustice by killing off humanity willy nilly. Thereby, it is only unfair to exhaust the energy pool in this way *if* our forebears have adopted safe-n-slow, but not dangerously fast.

Moving along, the two main policies my interlocutor is deciding among are both constrained by this conclusion regarding population size. Going safe-n-slow, bear in mind, means that they will always strive to bring further generations into existence (without dipping below the lower-bound). As I have said before, I will assume that they develop at half the speed of going fast. The goal is to stretch out humanity's tenure as

[1]This lower-bound might shift here and there depending on circumstances. So, for instance, if we are able to colonize the galaxy, then the the lower-bound will be greater for this and subsequent generations. After all, if it did not increase, then a small handful of persons would populate each planet; and this is less but still very risky than the Adam and Eve example provided.

long as is theoretically possible in the universe. This strategy is risky, however. It only takes one freak catastrophe to either annihilate humanity or bring about an especially severe broken world (from which they might not recover from in time, irreversibly lowering the average lifetime welfare of persons in this outcome). This could happen very early in human history. And if so, all of their sacrifices along the way would have been for nothing.

Compare this to the safe-n-slow policy. Adopting the safe strategy means that steps were taken by their forebears to protect their potential to recover from a freak catastrophe. So, while they cannot prevent a freak catastrophe, this horrific event need not spell doom for the population's long-term prospects. But how quickly they recover depends on the severity of the catastrophe, as well as which $\mathbb{E}$s were lost along the way—e.g., a mutated pathogen might destroy most of the world's water supply. Moreover, because they have taken all the right steps (regarding safety-engineering, counter-measures, and catastrophic risk reduction in general) we will say that an avoidable catastrophe is always successfully avoided by them.

By contrast, going dangerously fast banks on being able to pull the rip-chord before things get too hairy. Once an existential catastrophe looks like it may poke its ugly head out, this policy requires creating one final generation of sufficiently many persons such that they go extinct happily. How long they can pull this dangerous stunt off depends on how plagued their world is by catastrophic risk inititally, as well as how strongly their actions influence the cumulative risk of catastrophe. Furthermore, how much they might benefit from going fast is itself up for grabs. In other words, we do not know how quickly or slowly the population might evolve. There are four types of development paths we will run on the toy model. Going fast will pay off more or less depending on which rate of development is accurate of the real world. Similar comments, of course, apply to safe-n-slow as well.

A third policy choice which I have not yet introduced is a hybrdid of fast and slow. In a nutshell, if we were to adopt this policy, the population would start off minimizing catastrophic risk. However, at some later stage in history, when they have acquired a larger energy pool by (e.g.) colonizing the Local Supercluster, safety-first would diverge from safe-n-sow by imitating the dangerously fast policy. It would instruct the population to create a sufficiently large number of persons within a single generation. But whereas going fast pulls the trigger because their actions have resulted in remarkably high avoidable risk, safety-first prescribes exhausting the remaining energy pool for the reason that there's a non-negligible chance of a freak catastrophe resulting in either their extinction or a broken world. It is, in this sense, very cautious about prolonging humankind too long; once they've got enough chips, they cash out.

6.2 Unpacking the Basic Model

The toy model approximates the evolution of mankind in our universe. I have greatly simplified matters, and this is intended only as a rough approximation of the real world.

The long-term potential of the population is initially limited to Earth. If they make

poor decisions, then the population will die in the cradle so to speak. Alternatively, they might spread out across the stars and survive beyond the destruction of our planet when the Sun becomes a red giant. I will assume that there are a total of 10,000 generations that could ever exist in the universe. The energy pool provided by the Earth itself can host 1,000 of them. For simplicity, I assume that more energy can be added to the energy pool twice in post-Earth history. Upon colonizing the galactic neighbourhood, the population can grow as large as 4,000 generations (including the initial 1,000). Later, when they have devised more efficient energy extraction processes, they can grow as large as 10,000—their full potential.[2] Call these the three stages of humanity.

On the toy model, the policy of going fast never results in a population larger than 1,000 generations.[3] They never evolve past the first stage in other words. If a freak catastrophe does not wipe them out, then they will exhaust their energy pool either at the 250^{th}, 500^{th}, 750^{th}, or $1{,}000^{\text{th}}$ generation. Precisely which depends on how badly plagued their world is by catastrophic risk, and the rate at which cumulative risk grows by their dangerous promotion of $\mathbb{E}$s. For simplicity we simply randomize these four expiration dates.

I expect there to be roughly two freak catastrophes during humanity's tenure on Earth. This is by my best understanding, a very conservative assumption insofar as it makes my case for safe-n-slow harder to make. I think the actual probability is lower. Let's say that in the event of a freak catastrophe the chances of going extinct are $\frac{1}{4}$. This means that there is a $\frac{3}{4}$ chance that a broken world will obtain. If they have adopted going fast, then turbulence is set to zero and subsequent generations persist with lives just barely worth living ($\mathbb{B}$). But if they instead adopted safe-n-slow or safety-first, then the severity of the broken world is determined by a randomly generated value. I refer to this as 'turbulence' in the toy model. It ranges from 0.01 - 1. *Importantly, this means that, very counter-intuitively, there is a risk of leading a life not worth living only if the safe-n-slow policy is chosen.* Again, this is just part and parcel of me being shockingly charitable; it is *not* something I believe is true of the policy. At any rate, the lifetime welfare level of some generation is a product of their development function and turbulence.

Inititally I was tempted to design the toy model such that in stage three the risk

[2]See Ćirković and Radujkov (2001) and Sandberg et al. (forthcoming).

[3]Following [PM], even our near-immediate forebears would see no benefit in setting in motion the means of our escape off a dying planet. Our immediate forebears might—e.g., our grandparents may be better off knowing that they saved their grandchildren from doomsday. And perhaps even our great-great-grandparents will too. But the further back we go, the weaker the benefit to those forebears. Indeed, I certainly don't run across many persons that even know their great-granparent's name, let alone their story. Most of us just don't extend our compassion that far back or forth in time regarding blood lines. Moreover, this benefit of compassion must be traded off against other things they could do to benefit themselves. It hardly seems unreasonable to claim, as Professor Stephen Hawking once noted in a talk (note: I can no longer remember where it took place), that it will take a huge investment very well in advance of our great adventure into space in order to succesfully pull off space-colonization. My point here is that by the time some of our forebears have some interest in doing so, it will be far too late.

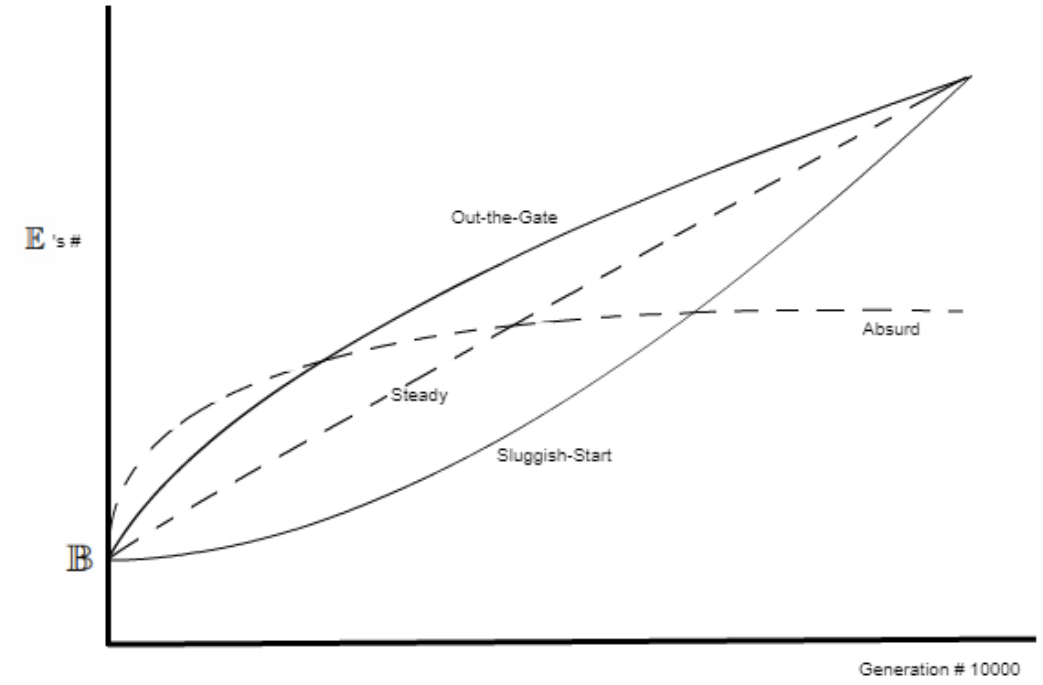


Figure 6.1: *4 Development Curves.* There are four development curves that I run on the toy model. They are: (a) out-the-gate; (b) absurd; (c) steady; and (d) sluggish-start.

of freak catastrophe will drop to a single freak event. Plus, I was going to assume that their development rate has switched from half-speed to full. This struck me then, as it does now, as appropriate. If the population has survived this long, then they are no longer as vulnerable to freak catastrophes (e.g., gamma-ray burst rarefaction), and will have defeated most threats. They won't have to take as many precautions at this stage of human history. However, I ultimately did not make these alterations as I was afraid of being called out for 'rigging the game'. Stage three is just like stages one and two in these respects.

On my toy model, I assume that our cave-dwelling ancestors started off with lives worth living. I have chosen to denote this with $\mathbb{B}$. You will notice that on my model there is no point at which humanity dips below a life worth living on the go-fast policy. Ignoring the charitability of having done so, perhaps this is a mistake on my part. Some of us might think that it would have been more interesting to see what came out if these policies could both result in lives not worth living.

Doubtful. Moreover, I truly struggle to imagine that life could get much worse than what our ancestors faced while hiding in cold, gloomy caves from ferocious creatures trying to eat them alive. This being said, I don't think tweaking the toy model to incorporate dips below $\mathbb{B}$ would produce any major differences to the conclusions reached here, apart from having go-fast lose even worse.

In the real world, we know that progress was sluggishly-slow at first before it skyrocketed. But even we do not know what's coming up next. It might be that progress peters out. By my best estimates there are roughly four types of development curves possible. These are illustrated in figure 6.1.

This is the simple model which we will use to generate a robust sample of outcomes for each policy. Below, I provide the script which was run on `Python` for all three policies. One script for each of the three policies—dangerously fast, safe-n-slow, and safety-first—run on the absurd development curve is provided here. (I call this the 'absurd curve' because it describes our earliest cave-dwelling ancestors as reaching approximately the same heights as we ever will. Things got good very, very early, and never really improved much. Whatever else we might think, this is decidedly absurd!) I provide the remaining three development curves separately afterwards. (Readers are encouraged to run the simulation for themselves.) The next section summarizes the results of running each policy under each development curve 1,000,000x in a table.[4]

```
import random
import math

n = 10000

PV_octopus = None

T = 1.0

def f_octopus(B, x, T):
```

[4]You will find, in the appendix, a different set of scripts, as well as their corresponding results, which are run on the Self-Sampling Assumption instead.

```
11
12      return (B + (4**4 * (math.log(x)) * T))
13
14
15  def run_model(B, L, T):
16
17      print('Running:        ', 'f_octopus (absurd takeoff)')
18
19      aggregate_average = 0.0
20
21      run_total = 0.0
22
23      size_total = 0.0
24
25
26      for y in range(1, 1000000000):
27
28          average = 0.0
29
30          B = 100.0
31
32          running_total = 0.0
33
34          L = 1000.0
35
36          D = random.randint(1, 4)
37
38          print('Doomsday occurs at ', L/D)
39
40
41
42          for x in range(1, n):
43
44                  freak_1 = random.randint(0, 500)
45
46                  survival_odds = random.randint(0, 4)
47
48                  #print('Freak Early: ', freak_1)
49
50                  #print('Survival odds: ', survival_odds)
51
52
53
54                  running_total = running_total + f_octopus(B, x, T)
55
56                  #print('Generation: ', x)
57
58                  #print('Lifetime welfare level: ', f_octopus(B, x, T))
59
60                  #print('          Running total: ', running_total)
61
62                  #print('Turbulence: ', T)
63
```

```
64
65
66              if (x > (L/D)):
67
68                      run_total = (run_total + (running_total +(f_octopus(B, x
    , T) * (L - x))))
69
70                      print('          Breaking out of model with running_total
    =', running_total)
71
72                      size_total = (size_total + x)
73
74                      print('          DOOMSDAY OCCURS AT:', x)
75
76                      #BLOCKED average = (running_total + (f_octopus(B, x, T)
    * (L - x))) / L
77
78                      #BLOCKED print('          Breaking out of model with
    running_total / x =', average)
79
80                      break
81
82
83              else:
84
85                      if (freak_1 > 499):
86
87                              if (survival_odds < 1):
88
89                                      print('          Breaking out of model with
    running_total =', running_total)
90
91                                      size_total = (size_total + x)
92
93                                      print('          DOOMSDAY OCCURS AT:', x)
94
95                                      #BLOCKED average = running_total / x if x !=
     0 else 0
96
97                                      #BLOCKED print('          Breaking out of
    model with running_total / x =', average)
98
99                                      break
100
101                             else:
102
103                                     T = 0.0
104
105                                     continue
106
107                     else:
108
109                             T = 1.0
```

```
110
111                         continue
112
113         print('
    RUN_TOTAL =', run_total)
114
115         print('
    SIZE_TOTAL =', size_total)
116
117
118         #BLOCKED aggregate_average = aggregate_average + average
119
120     #BLOCKED final_average = aggregate_average / y if y != 0 else 0
121
122     VAU = run_total / size_total
123
124     print('                    Final score: ', VAU)#BLOCKED final_average)
125
126
127
128 run_model(f_octopus, L=1000, T=T)
```

Listing 6.1: Toy Model: *Going Dangerously Fast*

```
1  import random
2  import math
3
4  n = 10000
5
6  PV_seppuku = None
7
8  T = 0.5
9
10 def f_seppuku(B, x, T):
11
12     return ((B + (4**4 * (math.log(x)))) * T)
13
14
15 def run_model(B, L, T):
16
17     print('Running:        ', 'f_seppuku (absurd takeoff)')
18
19     aggregate_average = 0.0
20
21     run_total = 0.0
22
23     size_total = 0.0
24
25
26     for y in range(1, 1000000000):
27
28         average = 0.0
29
30         B = 100.0
```

```

    running_total = 0.0

    L = 1000.0


    for x in range(1, n):

        freak_1 = random.randint(0, 500)

        freak_2 = random.randint(0, 2000)

        freak_3 = random.randint(0, 6000)

        survival_odds = random.randint(0, 4)

        severity_of_doomsday = random.random()

        #print('Freak Early: ', freak_1)

        #print('Freak Later: ', freak_2)

        #print('Freak Out: ', freak_3)

        #print('Survival odds: ', survival_odds)


        running_total = running_total + f_seppuku(B, x, T)

        #print('Generation: ', x)

        #print('Lifetime welfare level: ', f_seppuku(B, x, T))

        #print('         Running total: ', running_total)

        #print('Turbulence: ', T)


        if (x < L):

            if (freak_1 > 499):

                if (survival_odds < 1):

                    run_total = (run_total + running_total)

                    print('         Breaking out of model with
running_total =', running_total)

                    size_total = (size_total + x)
```

```
83
84                              print('          DOOMSDAY OCCURS AT:', x)
85
86                              #BLOCKED average = running_total / x if x !=
    0 else 0
87
88                              #BLOCKED print('          Breaking out of
    model with running_total / x =', average)
89
90                              break
91
92                        else:
93
94                              T = severity_of_doomsday - 0.5
95
96                              continue
97
98                  else:
99
100                       T = 0.5
101
102                       continue
103
104
105           elif (x < 4000):
106
107                 if (freak_2 > 1999):
108
109                       if (survival_odds < 1):
110
111                             run_total = (run_total + running_total)
112
113                             print('          Breaking out of model with
    running_total =', running_total)
114
115                             size_total = (size_total + x)
116
117                             print('          DOOMSDAY OCCURS AT:', x)
118
119                             #BLOCKED average = running_total /x if x !=
    0 else 0
120
121                             #BLOCKED print('          Breaking out of
    model with running_total / x =', average)
122
123                             break
124
125                       else:
126                             T= severity_of_doomsday -0.5
127
128
129                 else:
130
```

```
                        T = 0.5

                        continue


            else:

                  if (freak_3 > 5999):

                        if (survival_odds < 1):

                              run_total = (run_total + running_total)

                              print('          Breaking out of model with running_total =', running_total)

                              size_total = (size_total + x)

                              print('          DOOMSDAY OCCURS AT:', x)

                              #BLOCKED average = running_total / x if x != 0 else 0

                              #BLOCKED print('          Breaking out of model with running_total / x =', average)

                              break

                        else:
                              T= severity_of_doomsday  -0.5

                  else:

                        T = 0.5

                        continue

      print('
RUN_TOTAL =', run_total)

      print('
SIZE_TOTAL =', size_total)


      #BLOCKED aggregate_average = aggregate_average + average

      #print('                    Final score: ', aggregate_average)

#BLOCKED final_average = aggregate_average / y if y != 0 else 0

VAU = run_total / size_total

print('                    Final score: ', VAU)#BLOCKED final_average)
```

```

run_model(f_seppuku, L=1000, T=T)
```

Listing 6.2: Toy Model: *Safe-n-Slow*

```
import random
import math

n = 10000

PV_burn = None

T = 0.5

def f_burn(B, x, T):

    return ((B + (4**4 * (math.log(x)))) * T)


def run_model(B, L, T):

    print('Running:        ', 'f_burn_babymaker (absurd takeoff)')

    aggregate_average = 0.0

    run_total = 0.0

    size_total = 0.0


    for y in range(1, 1000000000):

        average = 0.0

        B = 100.0

        running_total = 0.0

        L = 1000.0


        for x in range(1, n):

                freak_1 = random.randint(0, 500)

                freak_2 = random.randint(0, 2000)

                freak_3 = random.randint(0, 6000)

                survival_odds = random.randint(0, 4)
```

```
48
49            severity_of_doomsday = random.random()
50
51            #print('Freak Early: ', freak_1)
52
53            #print('Freak Later: ', freak_2)
54
55            #print('Freak Out: ', freak_3)
56
57            #print('Survival odds: ', survival_odds)
58
59
60
61            running_total = running_total + f_burn(B, x, T)
62
63            #print('Generation: ', x)
64
65            #print('Lifetime welfare level: ', f_burn(B, x, T))
66
67            #print('         Running total: ', running_total)
68
69            #print('Turbulence: ', T)
70
71
72
73            if (x < L):
74
75                  if (freak_1 > 499):
76
77                        if (survival_odds < 1):
78
79                              run_total = (run_total + running_total)
80
81                              print('         Breaking out of model with
    running_total =', running_total)
82
83                              size_total = (size_total + x)
84
85                              print('         DOOMSDAY OCCURS AT:', x)
86
87                              #BLOCKED average = running_total / x if x !=
     0 else 0
88
89                              #BLOCKED print('         Breaking out of
    model with running_total / x =', average)
90
91                              break
92
93                        else:
94
95                              T = severity_of_doomsday -.05
96
97                              continue
```

```
98
99                  else:
100
101                         T = 0.5
102
103                         continue
104
105           #else:
106
107                 #run_total = (run_total + (running_total + (f_burn(B, x,
      T) * (4000 - x))))
108
109                 #BLOCKED #average = (running_total + (f_burn(B, x, T) *
    (4000 - x))) / 4000
110
111                 #BLOCKED #print('          Breaking out of model with
    running_total / x =', average)
112
113                 #break
114
115
116
117           elif (x < 4000):
118
119                  if (freak_2 > 1999):
120
121                         if (survival_odds < 1):
122
123                                run_total = (run_total + running_total)
124
125                                print('          Breaking out of model with
    running_total =', running_total)
126
127                                size_total = (size_total + x)
128
129                                print('          DOOMSDAY OCCURS AT: ', x)
130
131                               #BLOCKED average = running_total /x if x !=
    0 else 0
132
133                               #BLOCKED print('          Breaking out of
    model with running_total / x =', average)
134
135                                break
136
137                         else:
138
139                                T = severity_of_doomsday -0.5
140
141                                continue
142
143                 else:
144
```

```
                        T = 0.5

                        continue


            #else:

                #run_total = (run_total + (running_total + (f_burn(B, x, T) * (8000 - x))))

                #BLOCKED #average = (running_total + (f_burn(B, x, T) * (10000 - x))) / 10000

                #BLOCKED #print('          Breaking out of model with running_total / x =', average)

                #break


            elif (x < 8000):


                if (freak_3 > 5999):

                    if (survival_odds < 1):

                        run_total = (run_total + running_total)

                        print('          Breaking out of model with running_total =', running_total)

                        size_total = (size_total + x)

                        print('          DOOMSDAY OCCURS AT: ', x)

                        #BLOCKED average = running_total / x if x != 0 else 0

                        #BLOCKED print('          Breaking out of model with running_total / x =', average)

                        break

                    else:

                        T=severity_of_doomsday -0.5

                else:

                    T = 0.5

```

```
                        continue


            else:

                run_total = (run_total + (running_total + (f_burn(B, x, T) * (10000 - x))))

                #BLOCKED average = (running_total + (f_burn(B, x, T) * (10000 - x))) / 10000

                #BLOCKED print('          Breaking out of model with running_total / x =', average)

                break

    print('RUN_TOTAL =', run_total)

    print('SIZE_TOTAL =', size_total)


    #BLOCKED aggregate_average = aggregate_average + average

    #print('                    Final score: ', aggregate_average)

#BLOCKED final_average = aggregate_average / y if y != 0 else 0

VAU = run_total / size_total

print('                    Final score: ', VAU)#BLOCKED final_average)


run_model(f_burn, L=1000, T=T)
```

Listing 6.3: Toy Model: *Safety-First*

The remaining three development curves are provided below.

```
def f_seppuku(B, x, T):

    return ((B + (((4**(3.39) * (math.log(x))) + (x/(5/2)))) * T)
```

Listing 6.4: Out-the gate

```
def f_seppuku(B, x, T):

    return ((B + (x * (1/(4**(3.82))) * (math.sqrt(x)))) * T)
```

Listing 6.5: Sluggish Start

```

def f_seppuku(B, x, T):
```

```
3
4     return ((B + (x/2)) * T)
```

Listing 6.6: Steady

6.3 The Results

After running the toy model, I arrived at the following results (by taking the expected total utility divided by expected population size).

			Safety First		
Takeoffs	*Go Fast*	*Go Slow*	Early-Pull	Mid-Pull	Late-Pull
Absurd	2842	1209	16342	10618	6326
Out-the-Gate	1633	1033	19234	11732	3821
Slugglish-Start	262	516	1784	3889	1565
Steady	505	636	5228	5194	2441

Table 6.1: Results

The results speak for themselves for the most part. Go-fast outperforms safe-n-slow only on the Absurd and Out-the-Gate development curves. But rather than quibble about whether Sluggish-Start or Steady better represent our real life predicament, I will simply point out that safety-first wins in every category—and by a lot!

Interestingly, what we find is that in almost every case it is better to cash in our chips shortly after leaving Earth. The only exception is under Sluggish-Start where instead it would be better to do so after some extended period of time (i.e., a mid-pull).[5] But this shouldn't suggest that this is indeed what is best for the real world. Early-pull would be best if indeed we could increase the population or the average lifetime welfare of that sub-population in the way described by early-pull. We might not be able to do so. In reality, there are going to be a number of hurdles to doing so. For example, persons might be opposed to calling it quits at that point. And indeed it does seem like a shame to have finally colonized space, leaving Earth far behind, just to go out like an unimaginative punk. There are also potential dangers in overpopulating that I have pretty much ignored throughout this dissertation. These costs must be factored in. A final point I'd make is that it would be even more difficult to organize a mass extinction event of this sort when humanity is spread out across several colonies.

But even if we suppose that early-pull were not possible for whatever reason, it's crystal clear that surviving until even closer to the final curtain still outperforms go-fast—and, again, by a lot! So, we would just end up with something strongly resembling safe-n-slow anyhow.

[5]To clarify, an early-pull means the 1,000$^{\text{th}}$ generation goes extinct upon acquiring the energy for 4,000 generations. Mid-pull has the 4,000$^{\text{th}}$ generation do this with their energy for 8,000 generations. And late-pull has the 8,000$^{\text{th}}$ generation do this with an energy pool sufficient for 10,000 generations.

The Competing Claims View, I therefore conclude, requires us to minimize catastrophic risk. It would be worse if we let humanity go extinct on Earth. Humanity's place in the cosmos ought to be prolonged until we are at least at a stage where we could cash in our chips for a massive payday. We cannot know what will happen next. Hell, being safe might end with humanity in ruin. In the absence of a crystal ball, though, this is our best shot at maximizing expected overall goodness *qua* what is good for persons.

7

Conclusion

> "It's snowing still," said Eeyore gloomily.
> "So it is."
> "And freezing."
> "Is it?"
> "Yes," said Eeyore. "However," he said, brightening up a little, "we haven't had an earthquake lately."
>
> A. A. Milne, *The House at Pooh Corner*

Even the best laid plans of mice and men might end in doom and gloom. Our world can be a scary place. And though we will increasingly have the power to make life good, it could all go to Hell at any point.

There is pro tanto reason, I have said, to develop as quickly as possible in order to make life as good as possible for the present generation. This gambit increases the risk of an existential catastrophe occuring, but we cannot know with certainty that a freak catastrophe won't happen anyway. My goal in the dissertation has been to show that nonetheless we ought to resist this dark temptation. We ought to go safe-n-slow, and minimize as much as we can the risk of existential catastrophe in our world.

To be sure, going extinct sooner rather than later is not itself wrong according to the Competing Claims View so far as it is not worse for persons. However, this view nevertheless endorses a version of Nick Bostrom's Astronomical Waste Argument. Specifically, policies which threaten to bring about a broken world in which countless future people suffer in agony, and from which there is no chance of recovery, are worse. We ought to take onerous steps to avoid a broken world obtaining, and in the event of a freak catastrophe, that our civilization will be able to recover from Hell.

That a coherent and sensible moral theory holds that we ought to prevent a broken world isn't going to make the headlines. This is about as low of a bar as one can imagine when it comes to being morally decent. However, a far more interesting implication of the Competing Claims View germinates from this simple observation.

> *It makes things go better to prolong humanity if such onerous steps towards militating against catastrophic risk are taken.*

Indeed, this is definitely the most philosophically interesting implication of the (modified) Competing Claims View that I teased out in my thesis, especially given that we started off by saying that extinction does not in itself make the world go worse.

Along the way, I also argued that we have reasons of fairness which require us to avoid overpopulation. A generation becomes overpopulated, I have claimed, whenever subsequent generation(s) are placed in the trilemma of choosing between: (a) having to lower their own size in order to maintain an equal distribution going forward; (b) exhausting the remaining pool of energy by maximally propagating; or (c) leaving the next generation with an even harsher cut to make (in terms of (a)). If our world weren't plagued by catastrophic risk, there would be nothing unfair about humanity choosing to go extinct sooner rather than later. This final generation, no matter how large in size, would not satisfy my definition of being overpopulated. However, our world *is* plagued by catastrophic risk. It would be unfair to our forebears, I argued, to exhaust the energy pool willy nilly (only) if previous generations have endured some burdens to prolong humanity's place in the stars. And our claims against shouldering the heavy burdens of protecting humankind just don't have much weight—beyond how badly off we are in absolute terms—once we factor in the worthwhileness of these sacrifices.

Finally, I also tired to resolve one of the better-worn issues in population ethics. More precisely, I put forward a (mathematically) well-behaved axiology and bridge principle that satisfied the normative reading of $\mathbb{BN}$. I think we can probably do better. But I hope what I have shown here will at least have moved that conversation in the direction of getting clearer about the intuition of neutrality.

Appendix A

Discussion: Hooker's *Prevent Disaster Rule*

> "It's hard to be brave," said Piglet, sniffling slightly, "when you're only a Very small animal."
>
> A. A. Milne, *Winnie-the-Pooh*

Brad Hooker argues that the cost of inculcating in everyone the prevent disaster rule places a limit on its demandingness. My aim is to show that this is not true of existential risk reduction. However, this doesn't spell trouble for the reason that removing persistent global harms significantly improves our long-run chances of survival. We can expect things to get better, not worse, for our population.

A.1 Preliminaries

Hooker's preferred brand of rule-consequentialism holds that an act is wrong if and only if it is forbidden by the ideal code. This is the set of rules whose internalization by the overwhelming majority of everyone everywhere in each new generation has maximum expected value in terms of well-being (with some priority assigned to the worse-off).[1]

I will focus on the *prevent disaster requirement* in this article. It instructs us to sacrifice more than usual—and break the other rules if necessary—in order to promote the good when faced with calamity.[2] The merit of everyone having internalized such a rule is clear. In exceptional circumstances where a sufficiently large amount of expected value is at stake, it would be far better on the whole were people capable of enormous sustained sacrifice. As Hooker puts it, "[when] necessary to save the world, or even just some significant proportion of humanity, one may be required to make an extreme—even the ultimate—self-sacrifice".[3]

[1] (Hooker, 2000, 32)
[2] (Hooker, 2000, 98)
[3] (Hooker, 2000, 169)

A. DISCUSSION: HOOKER'S *PREVENT DISASTER RULE*

Here's the rub. The prevent disaster requirement threatens to make rule-consequentialism problematically overlydemanding. Take global hunger, for example. The starvation of innocent people surely counts as a disaster. However, an unlimited requirement to prevent disaster would entail, in so far as there is a staggering number of persons affected by food insecurity, that the relatively affluent should give up nearly everything, even forgoing personal projects or deep personal relationships—at least to the point where they are themselves impoverished.[4]

The reply from Hooker is that this criticism overlooks an important variable in our moral evaluations. The expected value of a rule is only partly determined by the benefits to the poor of getting aid and the costs to the affluent from donating. Missing from this equation are the costs of getting the rule internalized by *everyone*—including those who are least able to promote the good.[5] Even the poorest, most disadvantaged persons, after all, may find themselves in a position to save another's life.

Now, the safe assumption is that inculcating and sustaining in everyone the desire to abide by a rule will be more difficult the more demanding the rule is. Such costs militate in favour of less demanding rules (in general) for two reasons.[6] First, some logically possible codes are so overbearing, alienating or intricate that too few if any persons are capable of adhering to the rules, no matter how incredible the benefits promised by such a code.[7]

Second, the time, energy, attention and mental duress of internalizing a very demanding rule about preventing disaster could be ruinous.[8] And note that the benefits received by the global poor or hungry would themselves be (to some extent) counterbalanced by the costs of internalizing in them a completely impartial altruism towards others. Furthermore, internalizing the code is not a one-off cost, but rather something incurred with each new person added to our population—repeating indefinitely. "It is not as if infants begin with as much altruism as their parents eventually internalized ... [each] new generation would need to be transformed from beings concerned mainly with immediate gratification, personal comfort, and self-assertion to impartial beings willing to make virtually endless sacrifices for others."[9]

At first blush, the internalization costs of a very demanding altruism towards (among other things) abating global hunger would outweigh the benefits of improving the lives of the poorest in our population. If so, some point short of this level of sacrifice would instead produce the most well-being, given the costs.

I accept that the costs of internalizing a very demanding rule about preventing disaster could, given the state of a world,[10] produce less overall good than a moderately demanding rule. However, I will argue below that existential risk—that is, a risk that "threatens the premature extinction of Earth-originating intelligent life or the

[4](Hooker, 2000, 165)
[5](Hooker, 2000, 170-171)
[6](Hooker, 1995, 26)
[7](Mulgan, 2015, 111)
[8](Hooker, 2000, 166)
[9](Hooker, 2000, 166)
[10]See (Hooker, 2000, 172-173).

permanent and drastic destruction of its potential for desirable future development"[11]— does not support this being true of our world.

A.2 Prolonging Human History

My case rests on the astronomical waste produced by failing to prolong mankind's place in the world.[12] I will start by rehearsing a classic thought experiment from Derek Parfit. He writes,

> I believe if we destroy mankind, as we now can, this outcome will be much worse than most people think. Compare three outcomes:
>
> 1. Peace.
> 2. A nuclear war that kills 99% of the world's existing population.
> 3. A nuclear war that kills 100%.
>
> (2) would be worse than (1), and (3) would be worse than (2). Which is the greater of these two differences? Most people believe that the greater difference is between (1) and (2). I believe that the difference between (2) and (3) is very much greater. ... The Earth will remain habitable for at least another billion years. Civilization began only a few thousand years ago. If we do not destroy mankind, these few thousand years may be only a tiny fraction of the whole of civilized human history. The difference between (2) and (3) may thus be the difference between this tiny fraction and all of the rest of this history. If we compare this possible history to a day, what has occurred so far is only a fraction of a second.[13]

Certainly, nuclear war would bring about tremendous losses for all of those alive at the time. This alone provides good reason to take steps towards avoiding a nuclear disaster.[14] But perhaps the greatest loss occurs when the last of humanity is snuffed out by nuclear winter, given the whole of humanity's future follows them to the grave. Indeed, the vast reduction of the possible sum of happiness due to our extinction is literally astronomical, given the long-term potential of humanity. The destruction of mankind would be, as is often said by the likes of Parfit, Sidgwick, and others, the greatest of all conceivable crimes.[15] It constitutes what Bostrom calls an *existential catastrophe*.[16]

[11](Bostrom, 2012a, 15)

[12]An allusion to (Bostrom, 2003a).

[13](Parfit, 1984, 453-453). See also (Parfit, 2011, 616, 620).

[14]An anonymous referee flagged the question of why persons fail to think of an existential threat as being like an attack by a malicious enemy. This pressing consideration is too big to address adequately in the article. I leave it for another day.

[15](Parfit, 1984, 454). For a similar conclusion that does not rely on the Argument from Additional Lives, the interested reader may turn to (Kahane, 2014).

[16]For a detailed treatment of existential risk reduction see especially (Bostrom, 2012a).

This brings us to the point I now wish to press. Yes, Hooker is correct to alert us to the cost of internalizing in everyone the prevent disaster requirement. But we should likewise recognize that—on the assumption that additional lives worth living are of contributive value for our population[17]—there are goods, not just in improving the well-being of the poor or worse-off, but in prolonging humanity. My contention is that once we take stock of these goods in our moral evaluations, then a more demanding rule about preventing disaster (that is, our extinction) will produce higher expected value.

This can be demonstrated by first imagining a population of n persons. n stands for the complete population of mankind across time in this possible world. Let's say that their average welfare is $\mathbb{W}$ and that this already includes the cost of internalizing a moderately demanding code. Straightforwardly, the total sum of well-being in this population corresponds to $n \cdot \mathbb{W}$. As seems likely, let's also say that there are lots of catastrophic things that have some chance of happening to this population in the future. The total risk of doomsday occurring is not important for our purposes. All that is required here is that we accept that doing more to lower the chance of mankind's extinction thereby (a) increases the expected size of this population, and (b) increases the cost of internalizing the ideal code by everyone in this new population. Furthermore, m will stand for members of the population that may have existed had humanity only not gone extinct.[18]

Let φ be a finite positive number, $1 \geq \varphi \geq 0$, which captures the relation between higher internalization costs and size of the population m caused to exist. In short, the more persons caused to extinct by reducing extinction risk, the worse the quality of life of the (variable) population. We can then say that the total sum of well-being in the population which obtains is equal to $(\mathbb{W}-\varphi(\mathbb{W}))(n+\varphi(m))$. Notice that if $\varphi = 0$, then average welfare remains the same as before (and $\varphi(m) = 0$). Conversely, if $\varphi = 1$, then they are scraping by on muzak and potatoes, a life barely worth living (and $\varphi(m) > 0$). A more demanding prevent disaster requirement is of greater expected value whenever $\big((\mathbb{W} - \varphi(\mathbb{W}))(n + \varphi(m))\big) > n \cdot \mathbb{W}$. That is to say, there is some value of m such that a more demanding rule is better than a moderately demanding rule, namely $m > \frac{n}{1-\varphi}$.

[17] In other words, my argument will not move those readers endorsing the Procreative Asymmetry—roughly, the idea that there is no moral reason to bring a person into existence just because her life would be happy, but there would be reason to prevent a life from coming into existence if it were not worth living (McMahan, 1981); (McMahan, 2009). I find no purchase in the Procreate Asymmetry. For those who do, though, my case can be reworded by narrowing our definition of existential catastrophe to what Tim Mulgan dubs a *broken world*. A broken world allows us to ignore the *poor unlived masses of possible lives* in our moral evaluations while preserving the notion of catastrophes that would decimate the long-term potential of humanity. Mulgan describes "[this as] a place where resources are insufficient to meet everyone's basic needs, where a chaotic climate makes life precarious, where each generation is worse-off than the last, and where our affluent way of life is no longer an option" (Mulgan, 2015, 93). Just as with extinction, the loss of comparable goods, had things only gone smoother in history for actual persons, is astronomical. For the remainder of the article I will refer only to extinction events for ease of explication.

[18] Of course, the exact size of m depends on where we cap the long-term potential of humanity in the cosmos, as well as the risk of extinction in this world.

The far cosmic horizon of potential human history is undoubtedly large enough that the inequality holds for our world.[19] We can neatly summarize this counter-intuitive result as follows: *If what's supposed to be holding back the demandingness of the ideal code are the costs of internalization, then the goods promised by a long future for humanity will swamp those costs in our moral evaluations.*

Notice, the problem can be cast two different ways. First, the loss associated with humanity's premature extinction is so great that even if the probability of a catastrophic event is very low, an expected value calculation suggests that we should strive to prevent its possible occurrence. And yet, there is something deeply puzzling about ruining the lives of all actual persons for the sake of humanity eking out a longer stay in the universe.

Second, you may have realized that the above implication bears close resemblance to the dreaded Repugnant Conclusion. The Repugnant Conclusion states that for any population, all with a very high quality of life, there must be some larger imaginable population whose existence, all else being equal, would be better despite their lives being barely worth living.[20] The mistake, as countless critics have noted, is that quantity (that is, size of population) should not be able to compensate for a stark reduction to their average quality of life.

I'm inclined to agree that this looks worrisome. For some, if this were the end of the story, it would surely act as a reductio ad absurdum of the view. But this is not the full story.

A.3 An Indirect Approach to Lowering the Threat of Extinction

In setting out our earlier comparison of the two populations it was assumed that only costs go up, never benefits. That is to say, $\mathbb{W}$ was fixed and the total sum of goods went up merely because the size of the population grew, despite internalization costs reducing average quality of life. Colouring in the picture, this corresponds to the scenario where, all else being equal, existential threats are directly targeted. To illustrate, this could amount to putting a lot of resources towards asteroid deflection programmes.[21]

[19] To see this, let's make the conservative assumption that the potential exists for at least 10^{16} human lives of normal duration. If the risk of doomsday is $x, 1 > x > 0$, then a back-of-the-envelope calculation shows that improving our chances of survival by even a tiny fraction of a percent $(x - 0.00001)$ results in an expected $100,000,000,000$ additional lives for our population. This astronomically huge number surely tips the scales, not the cost of doing a little more to reduce existential risk by a tiny fraction of a percent.

[20] (Parfit, 1984, 388)

[21] As I am asking readers to understand the difference between direct and indirect approaches, one indirectly averts the threat of an existential catastrophe by solving intermediary problems which provide benefits outside this particular intervention. A direct approach to reducing the risk of nuclear war, for example, could involve sending UN inspectors to Iran to ensure regulations are being followed with respect to their nuclear facilities. This solution offers benefits isolated to the nuclear threat. On the other hand, sending aid packages to feed and clothe the poorest of Iran's people extends beyond helping the poor by (arguably) stabilizing geopolitics in that region of the world.

A. DISCUSSION: HOOKER'S *PREVENT DISASTER RULE*

I now wish to argue that we could instead reduce existential risk by indirect means, and in so doing make the world in two ways go better. As noted earlier, we would prolong humanity's place in the cosmos. Furthermore, an indirect approach improves the average welfare of persons, particularly the worse-off in our population.

Certainly, it would be a mistake to concentrate *exclusively* on indirectly lowering the probability of doomsday. Returning to our earlier example, reducing global poverty cannot prevent an Earthbound asteroid the size of Texas from making impact. Nevertheless, if we were *also* to adopt an indirect approach, then this would contribute to existential risk reduction by curbing the *negative ripple effects* of readily preventable illnesses, global hunger, and so forth.

Ripple effects are a class of phenomena that affect the far future in significant ways, shaping how our history unfolds over time.[22] A ripple effect is initiated by a particular event that has some causal influence on the course of events that follow it. These events, in turn, may have their own impact on how further events play out. And so on it goes, reaching wider and wider as time passes.

Consider the following example. A doctor is in a position to cure some infant's blindness. Sure, the infant will probably have a better life after the operation. Most of us are quick to hone-in on this feature of the situation. And many other goods go unacknowledged by us as a result. Just a few of the proximate advantages we might reasonably expect to find after curing the infant's blindness include: her parents will be less worried about her, subsequently finding more free time to develop their own personal projects; the government will spend fewer resources on providing her education; this child will grow up with more opportunities, as well as perhaps being inspired to start a grassroots initiative or develop an anti-malarial drug. All of these consequences will have some role in shaping our future due to their own ripple effects. This network of ripple effects might go so far as causing "[her] country's economy to develop very slightly more quickly, or make certain technological or cultural innovations arrive more quickly".[23]

My claim is that intervening in persistent global harms will contribute to existential risk reduction by removing harmful ripple effects in two important respects. I will focus on global hunger in making my case. My comments, however, are meant to apply generally—at least, to all cases of persistent global harms that might affect our leverage opportunities with respect to averting threats to humanity's future.

To begin, global hunger tends to make things worse around the world by its ripple effects, culminating in a range of independent doomsday scenarios. Indeed, not doing as much as is now possible to alleviate world hunger will mean that many more die from hunger and malnourishment-related illness. All of these deaths, consequently, produce their own growing burden in their communities. For example, a loss of working-hands in the fields means that their community's agricultural output suffers. An economy that depends on its agriculture sector could fall into disrepair and itself churn out

[22] See especially (Beckstead, 2013a). These are sometimes referred to as cascading, knock-on, or flow-through effects.

[23] (Beckstead, 2013a, 6)

further burdens as jobs and investments are lost. Consequently, the entire region could become socio-politically destabilized, fraught with civil war and marred by monstrous human rights violations. And this may introduce a hundred-year period of international hostility which invites the potential for nuclear engagement between warring parties. Or perhaps this region instead refuses to abide by treaties, such as by pumping out excessive greenhouse gases by mining coal in a fragile, protected rainforest.

Second, by failing to do more now to alleviate global harms our civilization is not just at growing risk of a catastrophe occurring, but also at growing risk of being unable to respond to an eventual existential threat. In other words, our leverage with respect to preventing a catastrophe could be severely weakened. It is not beyond the realm of possibility that in the aforementioned region, institutions equivalent to FEMA (that is, the Federal Emergency Management Agency) could become badly under-resourced, lacking the necessary funds and technical expertise to handle certain situations such as massive flooding—due to cancellation of training programmes, for example—or even non-existent if the government is corrupted by its military leaders, for example, and refuses to continue funding social programmes altogether.

I will now comment on some further benefits made possible by this approach.[24] Just as global hunger has ripple effects, so does the absence of starvation in this region. The primary benefit that springs to mind here is that global coordination is far less likely to degrade between well-fed, happy populations than it is with military dictators of oppressed states. And when it comes to existential risk reduction, our success in coordinating globally is of high importance.[25] Moreover, well-fed persons will be free to become educated, grow their economy, perhaps even getting involved in our civilization's scientific advancements, and so forth. They will make their children's future better. Lastly, if we solve global hunger, we will have, in effect, recruited more moral agents to aid us in eradicating malaria, poverty, or whatever happens to be next on our moral agenda. The altruistic burden will be spread out, and, perhaps after some extended period of incredible sustained sacrifice by our population, $\mathbb{W}$ will be higher than it has ever been in the past.

Tyrants and so on do not come from nowhere. They have a history. It seems to me clear that one plausible explanation for their presence in our world is that things tend to get worse around the world if (among other things) global hunger persists. In the hypothetical region we have been considering, if no one went hungry, then it is far more likely their crops would be tended, their cause for civil war and monstrous human rights violations jettisoned, and so forth. Therefore, the risk of humanity's extinction goes down by removing this fodder for catastrophe.[26] And, of course, abating global hunger

[24]The complexity of priority-setting according to ripple effects from persistent global harms is, to be sure, immense. It could (correctly) be said that such a rule is too complicated to be effectively internalized by persons. However, this shouldn't count as another cost of internalization since calculations of this sort should take place at an institutional (or collective) level, not around the dinner table. Rather, individuals have the burden of internalizing a general rule towards minimizing existential risk.

[25]See (Bostrom, 2012a, 21, 24-25).

[26]Hilary Greaves raised the worry (in personal communication) that speeding up our development in this way might present with its own existential risks (even if removing persistent global harms happens

means that $\mathbb{W}$ (in expectation) goes up. Taken in tandem, the conclusion we reach is that rule-consequentialism ought to prescribe a very demanding rule about existential risk reduction, and, by splitting our efforts between direct and indirect approaches, humanity's future could be very bright indeed.

A.4 Discussion

It might be said that my argument assumes that a mixed approach to existential risk reduction will make the world go better than a highly targeted approach would; *however* I have not shown this to be the case. In the absence of this further argument, rule-consequentialism may still be guilty of allowing impersonal concern towards prolonging humanity to swamp our moral considerations in ruinous ways. Yes, our population might last for a much longer time. But, the critic maintains, their living conditions would be bleak, the average life just barely worth living.

While not quite a storm in a teacup, this criticism misunderstands the aim of my argument. The pressing question, where the critic is quite right to demand an answer, is whether rule-consequentialism supports more (or most) emphasis being placed on highly targeted approaches.

Well, I have already mentioned this as a possibility. I will now add that our response to an existential threat depends on the circumstances in which we find ourselves. For example, supplying aid to the poorest among us cannot avert the threat from an Earth-bound asteroid the size of Texas. (But if this asteroid's trajectory were such that two hundred years will pass before its arrival, then we could, by concentrating on indirect interventions in the meantime, significantly improve our leverage opportunities with respect to destroying the threat.) Moreover, the solution to a particular threat (scheduled for fifty years from now) may be so difficult or time-consuming to figure out that allocating any resources to mitigating global hunger would seriously harm humanity's chance of survival.

Although this does not exhaust the range of reasons we might (regrettably) have for focusing exclusively on directly reducing the threat of extinction, they seem to

to limit or block certain pathways towards humanity's doom). To illustrate, we might put ourselves in the position of creating a dangerous technology sooner. My own reaction is that anthropogenic threats of this kind present as a step risk, not a state risk. That is to say: the severity of the risk depends on how we transition from an earlier to a later stage – not how long we are exposed to the possible threat. In Bostrom's words, "[the] amount of step risk associated with a transition is usually not a simple function of how long the transition takes. One does halve the risk of traversing a minefield by running twice as fast" (Bostrom, 2014, 234). As I see things, an indirect approach should be determined by considering how our leverage with respect to abating catastrophic risk is affected. As I have argued, eradicating global hunger will mean we are better prepared for handling existential threats. More so, I argued that allowing global hunger to continue ravaging our world will harm our chances of surviving such threats. So, we are better off even if the threat presents sooner than it would have had global hunger continued to plague the world's poorest. Therefore, it seems we can expect the threat of catastrophe won't go up *simply* by accelerating our development. However, that being said, we are problematically *clueless* about how particular interventions will shape the far future. For relevant discussion, see especially (Greaves, 2016).

share a common thread. It would be a mistake to allocate finite resources to indirect approaches whenever (a) an extinction event is looming and (b) capacity-building (or other indirect measures) would fail to impact our ability to prevent this particular catastrophe. But put this way, it is hard to see why the critic finds *this* response from rule-consequentialism objectionable. The call to arms seems appropriate when facing an impending apocalypse. And, indeed, my goal has never been to show that a mixed approach is (in principle) superior to a highly targeted approach. We *should* abandon indirect approaches when harrowing circumstances demand it.

A.5 Summary

I have argued here that the cost of inculcating in everyone a very demanding rule about minimizing existential risk (or curbing the astronomical waste produced by a broken world) does maximize expected value. My main thesis, however, was that this only looks problematic at first blush. As we have seen, taking an indirect approach to existential risk reduction, and so removing persistent global harms along the way, can improve the well-being of our population in significant ways. In the long run, we can expect things to get better, not worse, for our population.

So how bothered should we be about the severe costs of internalizing a very demanding impersonal concern for humanity's future? Does the conclusion we have reached count against rule-consequentialism? I do not see why it should. Hooker himself notes that this "possibility is a consequence of the idea that moral rules are to be selected from an impartial point of view. And this impartiality is a deeply attractive feature of rule consequentialism."[27] I wholeheartedly agree. And for those who still believe something has surely gone wrong here, I'll close by reminding them of the broken world that might otherwise lie ahead. As Mulgan puts it, "[can] we coherently picture the grim reality of their moral lives, and still refuse to make similar sacrifices for ourselves?".[28] Buck up.[29]

[27](Hooker, 2000, 172)

[28](Mulgan, 2015, 113)

[29]I wish to thank Mikio Agaki, Campbell Brown, Ben Colburn, James Humphries, Robyn Kath, Chris Mills, Japa Pallikkathayil and Catherine Robb for their insightful feedback on earlier drafts of the article. I am especially indebted to two anonymous referees at `Utilitas` for providing highly detailed comments that sharpened both my thesis and its presentation. For fruitful discussion, I am grateful to Nick Bostrom, John Cusbert, Daniel Dewey, Eric Drexler, Hilary Greaves, Daniel Kokatajlo, and participants of the *Ethics for a Broken World Conference* in Munich (especially Andrew Crabtree, Lisa Herzog, & Tim Mulgan). Finally, I extend my gratitude towards the Future of Humanity Institute for hosting me, as well as supplying the grist from which this article's arguments were formed.

A. DISCUSSION: HOOKER'S *PREVENT DISASTER RULE*

Appendix B

Expected Actualism, Dutch-Books, & Fortune-Tellers

> Pooh got in. He was just beginning to say that it was all right now, when he found that it wasn't, so after a short drink, which he didn't really want, he waded back to Christopher Robin.
>
> A. A. Milne, *Winnie-the-Pooh*

This short notice presents a criticism of a forthcoming view defended by Daniel Cohen, *Expected Actualism*. I produce two novel objections here: *Devil's Dutch-Book* and *Fortune-Teller's Admonition*. These are meant to show that, although Cohen's view is not subject to the same criticism as Moral Actualism, it nevertheless faces two residual problems. Firstly, Expected Actualism is a nonideal theory which only ideal agents can apply—so, is an exceptionally poor action-guiding theory. Secondly, Cohen's view implies that there is moral reason to bring mere additions into existence—so, he failed to achieve his goal of explaining the Procreative Asymmetry.

B.1 Prelims

For ease of explication, I'll begin by introducing some light notation, and stating one particularly important assumption.

Let's say that the set of all possible persons, $\mathcal{L}$, is denumerable, and number the people so that $\mathcal{L} = \{1, 2, ..., n\}$. We shall denote the zero well-being of person $i \in \mathcal{L}$ in an outcome where he does not exist with Ω. His life is worth living if he has positive (lifetime) welfare. If his welfare level were instead negative, then we consider his life to be not worth living. An important assumption made in the paper is that non-existence is *better for* a person than a life not worth living, and *worse for* the person than a life worth living.[1] Moreover, let $\Phi(\text{Wx})$ be an action, Φ, such that a world, Wx, would

[1] This position is commonly referred to as Existence Comparitivism in the field of population ethics.

obtain if Φ were performed. Finally, I'll hereafter write 'permissible$_{Wx}$' and 'good$_{Wx}$' as shorthand for, respectively, 'permissible relative to Wx' and 'good relative to Wx'.

B.2 Actualism, Criticism, & Cohen's Reply

According to *Moral Actualism*, what matters is how things go for actual persons—this includes our forebears, us, and all those persons that will be caused to exist. We hive off, in other words, all those possible persons that would have existed had things gone differently. The moral permissibility of an action depends only on how good it is for actual people. Assuming the total welfare of the population ought to be maximized, this gets us

> *Moral Actualism*: Φ(Wx) is morally permissible iff actual people have at least as much total welfare in Wx as they have in any other available world.

Important to bear in mind here is that permissibility isn't determined, on this view, by how an action affects those who would be actual if the action were performed.[2] Rather, permissibility is solely determined by how the action affects those who are in fact actual, given the action that is in fact performed.

B.2.1 Problem: *Choice-Dependence*

A well-worn problem for Actualism is that it's choice-dependent. This makes moral deliberation otiose; in order to know how we ought to act, we would already know how we will act. How else would we know which are the actual persons under consideration? This makes it a poor action-guiding theory insofar as it violates *normative invariance*: the normative status of an action does not vary depending on whether it is, or is not, actually performed.[3] Looking at the toy example, consider if Wx were the actual world. If so, then Jeff, an actual person, would have had higher lifetime welfare in Wy. Therefore, you ought not have brought him into existence. But if Wy were instead the case, then it would be morally permissible to bring Jeff into a horrible existence (Φ(Wx)).

World	Jeff	George
Wx	-10	Ω
Wy	Ω	Ω
Wz	Ω	-20

Table B.1: Ought Jeff Suffer?

Many consider this objection to be fatal.

[2] cf. (Hare, 2007)
[3] (Carlson, 1995, ch.6)

However, not everyone thinks a poor action-guiding theory is the end of the world. Krister Bykvist for one reminds us that many still-standing moral theories fail to provide guidance on every occasion. Take, for instance, consequentialism, which holds that the consequences of an act are all that matter. Yet, "lacking a crystal ball, how could you possibly tell what *all* the effects of your act will be? So how can we tell which act will lead to the best results overall—counting *all* the results?"[4] The theory on offer, Bykvist goes on to argue, is only in deep trouble—whether or not it violates normative invariance—if it fails to be *satisfiable*. He proposes the following requirement be met for a theory to have cut mustard.

> *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.[5]

Consequentialism is satisfiable because an ideal observer could correctly apply the theory. By contrast, Actualism creates dilemmas where there is no way out. There are scenarios where even the ideal agent will regret whichever action he ultimately chooses to perform.[6] Comparing only Wx and Wz in our toy example, if you create Jeff, he will lead a horrible life, and so you ought to Φ(Wz). But if you instead create George, then you ought to have performed Φ(Wx). Therefore, no matter which horrible life you bring into existence, you do wrong. This isn't however your traditional dilemma where every alternative is wrong. Rather, no matter how you act, the other action turns out being the morally correct choice for you to have made.[7] In Bykvist's words, "[since] these implications hold no matter how ideal we assume we are, there is reason to think that this is a failing in the theories rather than in us."[8]

B.2.2 Solution: *Maximize Expected Actual-World Permissibility*

Cohen proposes to salvage *Moral Actualism* from the wreckage by supplementing it with a theory of subjective permissibility.[9] If one doesn't know which world is actual, then he must resort to a subjective decision procedure which maximizes *expected* actual-world permissibility. Let's call the view that Cohen develops *Expected Actualism*. According to him, Expected Actualism is (a) action-guiding; (b) respects normative invariance; and (c) explains the Procreative Asymmetry; roughly, the position that we ought not cause a person to exist if they were to be miserable, but we do not have corresponding moral reason to bring a happy person into existence *just because* she would be happy. Below, I'll begin by outlining Cohen's theory, and showing how it does not get gored on the same horns as Actualism.

[4](Kagan, 1998, 64)

[5](Bykvist, 2007a, 116)

[6]See (Bykvist, 2007a, 112, 116-118). Bykvist refers to Actualism throughout as the 'world-relative preference-affecting theory'.

[7]Actualism does not factor in for the fact Jeff would suffer less over his lifetime than George.

[8](Bykvist, 2007a, 118)

[9](Cohen, forthcoming)

B. *EXPECTED ACTUALISM, DUTCH-BOOKS, & FORTUNE-TELLERS*

The first step in Cohen's framework is to determine permissibility facts for each possible world W, as determined by how those in W fare in different worlds. In terms of axiology, Wx is considered to be at least as good_{Wy} as Wz iff the Wy people have at least as much total welfare in Wx as they have in Wz. Bear in mind, each W will generate its own world-relative permissibility facts. We are governed by those permissibility facts relativized to the actual world. Of course, we cannot always know, antecedently, which world is actual, and, so, which actions are objectively permissible. However, we do have access to whether $\Phi(\text{Wx})$ is $\text{permissible}_{\text{Wx}}$ and so on for each Wy. Cohen suggests that, in light of our deep ignorance in such circumstances, we ought to *maximize expected objective permissibility*. He proposes that

> $\Phi(\text{Wx})$ is *subjectively permissible* iff its degree of $\text{permissibility}_{\text{Wx}}$ is at least as great as the degree of $\text{permissibility}_{\text{Wy}}$ of $\Phi(\text{Wy})$, for every available $\Phi(\text{Wy})$.[10]

The degree of $\text{permissibility}_{\text{Wx}}$ of $\Phi(\text{Wx})$ is equal to the welfare of the Wx people in Wx minus the welfare of the Wx people in Wy, where Wy is an available world in which the Wx people have greater welfare than they do in any other available world.[11,12] The 'permissibility value' of all permissible actions is 0 because the difference between the highest welfare outcome and itself (or any outcome with the same amount of welfare) is, of course, 0.[13] There might be several outcomes for which there's a score of 0. Actions that are impermissible relative to the world in which they are performed, however, can present with different values below 0. So, following Cohen's model, we are able to rank different impermissible actions from worse to worst.

Right off the bat, we can see that Cohen's view isn't choice-dependent. So, it won't violate normative invariance. This is good news. And there are, I think, four more instructive points worth spelling out carefully here regarding Expected Actualism's prescriptions in population ethics. See the tables below.

World	Jeff	George	James	EP
Wu	-10	Ω	Ω	-10
Wv	Ω	-20	Ω	-20

Table B.2: Expected Actualism in Action: Problem Set 1

Let's start with problem set 1. The expected $\text{permissibility}_{\text{Wu}}$ of $\Phi(\text{Wu})$ = Jeff's welfare in Wu (-10) minus the greatest welfare achieved in an available world (Ω). So, just -10. Performing the same operation for George produces an expected permissibility score of -20. Therefore, though both are impermissible actions, causing Jeff to exist is less impermissible (in expectation) than causing George to exist. This is intuitively plausible.

[10](Cohen, forthcoming, 7)
[11](Cohen, forthcoming, 9)
[12]Cohen ascribes the basic idea to a consideration taken up in (Wedgwood, 2011).
[13](Cohen, forthcoming, 8)

World	Jeff	George	James	EP
Ww	10	15	Ω	0
Wx	10	10	Ω	-5
Wy	10	Ω	Ω	0
Wz	10	Ω	20	0

Table B.3: Expected Actualism in Action: Problem Set 2

To flesh out the next three considerations, let's turn to problem set 2, and imagine that Jeff is currently alive. George and James would be mere additions—that is to say, bringing either into existence will in no way affect the original population (Jeff). George could have 15 or 10 lifetime welfare depending on whether Jeff performs Φ(Ww) or Φ(Wx), respectively. An alternative for Jeff is Φ(Wy), in which case George never comes into existence. As Cohen points out, Expected Actualism gets the correct result in that if the same persons would exist in multiple outcomes, we ought to realize the world in which those persons are best off. To see this, notice the expected permissibility$_{\text{Ww}}$ of Φ(Ww) = Jeff and George's welfare in Ww (25) minus the greatest welfare achieved by them in an available world. In this case that world happens to be the same, Ww; therefore, the score generated is 0. In Wx the total welfare of actual persons is reduced by 5. So, subtracting 25 from 20 produces a score of -5. So far so good. A third item worth highlighting is that the expected permissibility$_{\text{Wy}}$ of Φ(Wy) = Jeff's welfare in Wy (10) minus the greatest welfare achieved by him in an available world, which is 10 in every outcome. Therefore, Φ(Wy) generates a score of 0, and is equally permissible with Φ(Ww). Therefore, we do not have moral reason to bring George into existence *just because he would be happy.*

A fourth consideration may, however, count against Expected Actualism. Given a choice between Φ(Ww) and Φ(Wz), where James is the happier of the two persons, there is no subjective obligation to bring him about over George.[14] Both actions have an expected permissibility score of 0. I won't explore this implication further, however, because person-affecting theorists disagree amongst each other as to whether this is the wrong result. Instead, I'll present two objections which everyone can agree are *prima facie* problematic.

B.3 2 Objections

To begin, observe that while Expected Actualism is choice-independent, it is nonetheless *choice-set-dependent*; whether or not Φ(Wx) ≻ Φ(Wy) can depend on which other alternatives, besides Wx and Wy, are in the choice set. More specifically, Cohen's view doesn't satisfy the conditions commonly known as *contraction consistency* and *expansion consistency*, which hold that the ranking of some outcomes ought not change

[14] As Cohen is fully aware, discussing it at (Cohen, forthcoming, 10).

when other alternatives are added or taken away from the choice set.[15]

World	Jeff	George	James
Wx	10	4	Ω
Wy	10	5	3
Wz	10	Ω	10

Table B.4: Reversing Rankings

In the above table, if we compare only Wx and Wy, Cohen's view implies that $\Phi(\text{Wy}) \succ \Phi(\text{Wx})$ for the reason that the expected permissibility of $\Phi(\text{Wx}) = -1$, while $\Phi(\text{Wy}) = 0$. However, if we were to add Wz to the choice set, then $\Phi(\text{Wz}) \succ \Phi(\text{Wx}) \succ \Phi(\text{Wy})$ (because the expected permissibility of $\Phi(\text{Wx}) = -1$, $\Phi(\text{Wy}) = -2$, and $\Phi(\text{Wz}) = 0$). The initial ranking of $\Phi(\text{Wy})$ and $\Phi(\text{Wx})$ has reversed.

There are, so far as I can see, at least two ways in which this feature undermines Cohen's proposed theory. I'll run through both problem-cases first. Afterwards, I will consider what I take to be the strongest counter-response from Cohen. As I shall argue, though the blow can be softened, the theory left standing lacks much of its initial appeal.

World	Jeff	George	James	EP
Wx	1	-1	Ω	-1
Wy	Ω	1	-1	0
Wz	-1	Ω	1	0

Table B.5: The Devil's Hellish Offer. EP Ranking: $\text{Wy} \succ \text{Wx}$, $\text{Wz} \succ \text{Wy}$, $\text{Wx} \succ \text{Wz}$

> *Devil's Dutch-Book.* Imagine the Devil offers you the chance to select the next population he brings into existence. For the price of a momentary stay in Hell you can pick between Wx and Wy. Let's suppose that you strongly prefer maximizing expected actual-world permissibility over spending a moment in Hell. After you make your choice, the Devil sneakily introduces another option. He repeats the offer, this time leaving you to chooose between

[15] According to the "*Independence of Irrelevant Alternatives*: For any two outcomes, A and B, to know how A compares to B all things considered it is , at least in principle, sufficient to compare them directly in terms of each of the ideals about which we care. More particularly, if one accurately knew how A compared to B in terms of each ideal relevant to our all-things-considered judgments, and if one granted each ideal its due weight, then one would be in a position to know how A compared to B all things considered" (Temkin, 2012, 387-388). This captures both contraction consistency and expansion consistency (of an option set). That we know how a population fares all-things-considered to another population is supposed to be guaranteed by the *Internal Aspects View of Outcome Goodness* (Temkin, 2012, 370). Together with *completeness*, this entails the *Principle of Like Comparability of Equivalents*: "if two outcomes or prospects are equivalent (meaning equally good) in some respect, then however the first of those outcomes or prospects compares to a third outcome or prospect in that respect, that is how the second of those outcomes or prospects compares to the third outcome or prospect in that respect" (Temkin, 2012, 237).

keeping your previous choice or switching to Wz at the cost of yet another moment spent in Hell. The Devil continues this for the rest of your stay in Hell.

As demonstrated in table B.4, so long as the Devil is in charge of what your choice set contains, you'll get dutch-booked for an eternity in Hell. This is clearly catastrophic. Consider now the following parable.

> *Fortune-Teller's Admonition.* Imagine a fortune-teller presents you with a button which whever pressed brings a happy person into existence somewhere in a disconnected part of the galaxy. This person would be a mere addition. (Φ(Wy) is pressing. Φ(Wx) is not.) You quickly surmise that you are permitted to perform either action in accordance with Expected Actualism. The fortune-teller, however, has peered into his crystal ball in the meanwhile. Sullenly, the fortune-teller declares that he sees a child in one of your future timelines which suffers from a horrible disease. The child's life will be, on balance, just barely not worth living. Yet, this child will enrich your life, even going so far as to make you a better person. Do you press the button now?

World	Jeff	George	James	EP_1	EP_2
Wx	10	Ω	Ω	0	-1
Wy	10	5	Ω	0	0
Wz	11	Ω	-2		-1

Table B.6: Mere Additions: Surreptitious Fortune-Telling

As is demonstrated in table B.4, the only permissible action left for you to perform is mere addition, Φ(Wy). Generally, then, for any conceivable mere addition case, I can approach the agent and *make it the case* they ought to press the button. Admittedly, this won't be for the reason that this person's happiness itself provides moral reason to cause them to exist. No, it would be for the very strange reason that I merely reported the possibility of *some* world in which you would have a child which overall has a life not worth living, who nevertheless improves your own life. Plainly, this isn't in the spirit of the Procreative Asymmetry. Those defending the Procreative Asymmetry, after all, wish to maintain that both options (Wx and Wy) are equally permissible.

B.4 Discussion

I can anticipate the following reply from Cohen. The Devil's Dutch-Book only goes through if the Devil can appeal to a collection of disparate choice sets. If all of the possible outcomes were contained within a single choice set, then we can scotch his dutch-booking. Indeed, by only being able to turn to a complete choice set, there will

never be a situation in which Expected Actualism gets money-pumped or reverses its ranking of the alternatives. Returning to *The Devil's Hellish Offer*, we see that if all three outcomes are simultaneously considered, all three are equally permissible (sharing a score of -1).

World	Jeff	George	James	EP
Wx	1	-1	Ω	-1
Wy	Ω	1	-1	-1
Wz	-1	Ω	1	-1

Table B.7: Scotching the Devil's Dutch-Book. EP Ranking: Wy $\succeq$ Wx, Wz $\succeq$ Wy, Wx $\succeq$ Wz

At first blush, this seems to be the correct reaction. However, it does nothing to overturn the implication of required mere addition in *Fortune-Teller's Admonition*. As already captured in table B.4, the expected permissibility score of mere addition can always be rigged by the inclusion of a possible world where (a) members of the original population are better off, but (b) it would be, on balance, better not to bring about this world (because, e.g., a person caused to exist would have a life not worth living). It doesn't, in other words, hang on whether an alternative shifts the ranking of outcomes by Expected Actualism.

Moreover, besides not explaining the Procreative Asymmetry, you may recall that Cohen put forth his theory as a nonideal alternative to Moral Actualism. It's what we turn to when we don't know which world is in fact actual. But it's impossible for us to know every alternative in the complete choice set required of Expected Actualism. The full list of possible outcomes is, no doubt, countably infinite. This is especially problematic, I submit. But not merely because it's a poor action-guiding theory—after all, Expected Actualism can meet Bykvist's condition of *satisfiability*. Rather, my central worry is that, unlike (*inter alia*) consequentialism, which could, for example, nonidealize their approach (e.g., temporal-discounting, sophisticated principles of indifference, or complex expected value calculations) in the face of *cluelessness*,[16] the theory on offer from Cohen was intentionally designed as a nonideal solution. But it's a nonideal solution available only to ideal agents. So, I wonder, where should we go next in search of practical guidance from Moral Actualism?

The combination of these two criticisms, I think, speaks strongly to the implausibilty of the view. Indeed, I can see no easy answers here for defenders of Expected Actualism.[17]

[16] See especially (Greaves, 2016); cf. (Lenman, 2000).

[17] The paper before you germinated from an illuminating conversation with Daniel Cohen and Michael Plant. I wish to thank them both for comments. *Note:* Cohen has recently modified his view in an attempt to get around my objections—this is to say, these alterations took place after my viva. I have chosen to keep this paper in the appendix as it is nevertheless instructive for the unfamiliar reader. At any rate, it would seem to have been in vain as his new view now finds itself gored on some of the problems that I presented in 5.5.2 instead.

Appendix C

The $\mathcal{OP}$ is No Place for Infinite Ethics

> "Isn't that fine?" shouted Winnie the Pooh down to you. "What do I look like?"
> "You look like a bear holding on to a balloon," you said.
> "Not," said Pooh anxiously, "not like a small black cloud in a blue sky?"
> "Not very much."
>
> A. A. Milne, *Winnie-the-Pooh*

The *Final Anthropic Principle*, proposed by Barrow and Tipler, states that intelligent information-processing must come into existence in the universe, and, once it comes into existence, it will never die out.[1] The claim that life must come into existence is, clearly, far too strong. We can remove this teleological element, weakening the hypothesis in two ways. First, we might maintain that once a race of intelligent information processing creatures comes into existence, it will never die out. Second, it could be said that any particular intelligent race might die out, but intelligent life as a whole will continue existing indefinitely.[2]

What we know about the world doesn't support either life itself or a given population surviving for eternity in the cosmos. Primarily, the presence of a non-zero vacuum energy density makes indefinite survival within a single cosmological domain impossible. And beyond the explosive shrinking of particle horizons, there is the further difficulty of depleting the energy necessary for information processing within the horizon of any observer (and, on a larger scale, the problem of *heat death*).[3]

[1] See (Barrow and Tipler, 1968).

[2] These two formulations of the Final Anthropic Hypothesis are presented in (Ćirković and Bostrom, 2000, 676).

[3] "One cannot do anything useful with the vacuum energy. As far as the [cold dark matter] is concerned, it could conceivably be used as an energy source, since the annihilation of these cosmions and anticosmions (present in approximately equal numbers according to the standard theory) would produce potentially usable energy. However, depending on the mass spectrum of cosmions, their galactic density is rather small, and since their interactions are by definition very weak, their gathering for

C. THE $\mathcal{OP}$ IS NO PLACE FOR INFINITE ETHICS

If these individual domains were but infinitesimally small pockets of the multiverse, however, over which quantum fields vary in a chaotic manner,[4] and if it were possible to travel from one domain to another, then perhaps the hypothesis could be salvaged. However, there are excellent reasons for assigning a tiny credence in this being true of our world. Primarily, why has our universe not been visited by some alien supercivilization originating from a bubble much older than our own? Probabilistically speaking, they should have done. As Bostrom & Ćirković put it, "[the] simplest solution is to assume that interbubble migration is impossible (which might be supported by independent physical evidence in due course)".[5,6] For another, as Bostrom and Ćirković go

annihilation will pose huge engineering problems. If we consider the model of "posthumanity soon", than the only usable matter field is the baryonic matter, in the local environment concentrated in the form of planetary systems (Ćirković and Radujkov, 2001, 54)."

[4]Things might be far wierder than we ever could have imagined. See (Barrow, 2012, 194ff).

[5](Ćirković and Bostrom, 2000, 684); cf. (Davies, 1978)

[6]There are many other possibilities of the same ilk that are guilty of Bostrom's objection, such as resupplying our energy pool by bringing it, by means of wormholes, from distant points in the universe or from some alien-galaxy within a thermalized region of a bubble of inflationary phase. For a technical report on the subject of using wormholes for such purposes, see especially (Garriga et al., 2000). After all, one must still wonder why some older, alien hypercivilization in the vastness of our infinte universe has not already tampered with our energy pool. Conversely, imagine that it were possible to create a baby-universe in the lab, into which a population could then migrate. Well, if that were so, then if the original universe is infinite and contains at least one hypercivilization capable of creating baby-universes, then we should expect there to be very many such baby-universes (because, for example, it makes sense for this mature hypercivilization to repeat the process every time the finite energy within a horizon of the baby-universe drops to a level were they would soon starve), far outnumbering the one original universe. Therefore, it's very probable that our own world is such a baby-universe sitting in a proverbial goo-filled jar on some shelf somewhere. So, its highly probable that an older, alien hypercivilization will have migrated through into our own universe. Yet, none of these ancient beings have yet to be found. It's possible to salvage the proposal on offer if the caveat is added that it's physically impossible for lower-level populations to come through, and so they instead recreate an approximation of their own population in the baby-universe (by shaping the final conditions of the baby-universe or by creating very many such baby-universes in random conditions on the chance that one or more contain life as they know it) in order to (e.g.) maximize happiness in the multiverse. (For more on the subject of growing baby-universes, see (E. Farhi and Guven, 1990) and references within.) Because the baby-universe population's existence is owed to our actions this makes them a sub-population of our own (eternal) population by my own choice of reference class. So, in such circumstances, the (eternal) population is not necessarily doomed. Nevertheless, we ought to ignore the sub-population belonging to that baby-universe and any further iterations of baby-universes within baby-universes. That's because our causal influence on them ends after bringing about conditions—creating a baby-universe in a goo-filled jar—underpinning their existence, and there's nothing further that our actions can do that affect their population trajectory. Furthermore, we can never know what their world would look like. We are utterly in the dark regarding their way of life. Their laws of physics might be different, it could be hellish or heavenly, and so on. Because we cannot discriminate amongst these possibilities, we can ignore them in our deliberations, and make our decision based on what we do know (or are merely *weakly clueless* about). But suppose their universe is just like our own. Even if that were so, we can ignore the baby-universe population trajectory in our deliberations for the reason that the overall value of this outcome is just the same according to VA. So here's my reaction to the baby-universe scenario. In our ordinary evaluations we should restrict our deliberations to just those members of the population that would exist in this universe *unless* migration is possible. (Bear in mind, should we ever find ourselves capable of creating baby-universes for such a purposes, the genesis of our world is, probabilistically-speaking, surely owed to some astronomically-ancient population from

on to say, in order to find purchase in the Final Anthropic Hypothesis, we must suppose the multiverse has a very particular structure, that migration between causally connected domains of the multiverse is physically permitted, and that the fraction of all domains that are inhabited by a supercivilization originating from some other domain will tend to be negligible (such that, it were consistent with the observation that our universe does not seem to have been colonized by any such older alien creatures).[7] These (minimally) three stringent requirements strongly deflate the credence we should have in the Final Anthropic Hypothesis being true.

My arguments in the dissertation were run on the assumption that life must come to an end. I assumed the truth of this proposition because we just don't have strong enough reason at this time to think that the final curtain isn't going to be pulled on life at some point in cosmic history.

But here, in the appendix, I'm willing to explore what would fall out of the Rawlsian $\mathcal{OP}$ if instead the interlocutor took the possibility of eternal life seriously. In a nutshell, the $\mathcal{OP}$ stops returning prescriptions which are consistent with our intuitions concerning what we owe each other as a matter of fairness. This is unsurprising. Infinities tend to mess everything up, and the subject of ethics is no rare exception in this regard.[8] At any rate, here are just some of the problems generated by taking seriously in the $\mathcal{OP}$ the possibility of a civilization surviving for an infinite amount of time.

To begin, the possibility of interbubble travel presupposes that this civilization is extremely advanced—at the very least, we must asssume they are very competent in manipulating the world around them (and finding energy in the darkest, most foreboding places). More so, there are very many very difficult challenges ahead for Earth-originating intelligent life that threaten to toss us into the dustbin of cosmic history. Yet, the Final Anthropic Hypothesis strongly suggests that (a) it will be very easy to transition from cave-dwelling Earthlings to absolute masters of our supercluster of galaxies and, worse, (b) that, given that our pathetic existence as cave men (and subsequently as victims of natural and man-made catastrophes (e.g., global hunger)) is but a tiny (finite) fraction of this population's eternal existence, there is little to no chance of not finding oneself, upon awakening from the $\mathcal{OP}$, in the golden era of this population. This is true even if we were to accept an even weaker formulation maintaining that once life comes into existence, it *could*, given the right steps are taken and right conditions met, exist indefinitely. After all, the tiniest credence in the existence of one or more infinite population trajectories will swamp the other possibilities, distorting their deliberations in predictably bad ways.

Consider trying to rank the following: (a) eternal population trajectories that are horrible for the finite, earlier sub-population but very good for the infinite, later sub-population; and (b) astronomically huge (but finite) population trajectories that go asymptote to the very good (at the same epoch as in (a), let's say) but do not contain the same kinds of horrors as in (a). It's easy to see how the eternal population in which

a lower-level universe! See especially (Bostrom, 2003b).)

[7](Ćirković and Bostrom, 2000, 685)

[8]See especially (Bostrom, 2011).

the forebears suffer immensely is preferrable from the perspective of the interlocutor (indeed, no matter how horrifying the early stages of their cosmic history prove to be)—after all, the interlocutors will consider the prospect of finding themselves in those early stages upon leaving the rabbit hole to be bordering on the probabilistically-impossible in (a), but not (b).

Nor does it matter how long the hellish existence of the earliest sub-population last. Imagine, to borrow the gedanken experiment from James Cain, a bubble of hell spreading out infinitely in every direction (where persons, let's say, are spread out uniformly in the world).[9] At any time we pause the scenario, there will be a finite amount of persons in the bubble of hell and an infinite number of them outside the bubble. If we think of this bubble as how long our civilization must endure hell before things become very good for their progeny, then it's clear that we can continue expanding the bubble indefinitely—without overturning the interlocutor's preference for (a), given the expected value of (a) will always be approaching the limit on very good (unlike (b))—so long as the sub-population outside the bubble is of infinite size.[10]

It's plain that the $\mathcal{OP}$ is ill-suited for addressing the remarkably unlikely scenario of an infinite population. Permitting this scenario under the umbrella of closest possible worlds will drive the interlocutor's deliberations away from those matters which, well... matter. To this end, I think we should exclude infinitely-big populations from the list of closest possible worlds. Partly this is because, as we discussed above, it's reasonable to assign a much higher credence in the actual population being doomed. Sooner or later Death *is* coming to harvest what is owed to him by right.[11] But there's also the little matter of what the $\mathcal{OP}$ can and cannot do. You will recall from chapter 2, I suggested that the hypothetical contract proposed by Rawls could not be applied to some parts of morality, namely population ethics. But whilst I do think—and this dissertation is proof, I hope—that population ethics can be salavaged in this respect, it's clear from my above arguments that infinite ethics is not such an area of morality to which the $\mathcal{OP}$ can be applied. At the very least, it doesn't seem as if we are going to generate any prescriptions consistent with our intuitive notion of fairness.

*
* *

My critics could reply that similarly-poor decisions will pour out of my suggestion for revamping Rawls' $\mathcal{OP}$ if the assumption of our guaranteed extinction someday is coupled with the (plausible) assumption that our population begins small, and each successive sub-population grows larger than the previous one. Their worry can be illustrated with John Leslie's gedanken experiment.

Demonic Shooting Room. Imagine that the Devil creates people in a room,

[9](Cain, 1995)

[10]A good starting place for those interested in unpacking the wickedly complex field of infinite ethics is (Bostrom, 2011). Other good sources include (Cain, 1995); (Arntzenius, 2014); and (Mulgan, 2002).

[11]The (good) question of whether going extinct sooner rather than later is morally bad is pursued in (Lenman, 2002).

> in a batch or several batches. Successive batches would be of 10 people, then 100, then 1,000, then 10,000, then 100,000, and so forth: each batch ten times as large as its predecessor. You find yourself in one such batch. At every stage in the experiment, whether there will be any later batches is decided by the Devil's two dice, thrown together. The people in any batch watch the dice as they fall. So long as they don't land double-six, the batch will exit from the room safely. If they fall double-six everybody in the batch is to be shot, the experiment then ending. How confident should you be of leaving the room safely?[12]

Stuck in his solitary, dark room, our poor interlocutor might become convinced that he will find himself in the end times of his population. After all, doomsday awaits 90% of his fellow interlocutors. This is bad news insofar as it promises to generate equally strange recommendations behind the veil of ignorance as does factoring in an infinite population.

So far as I can tell, there are at least two counter-replies available to me here. Firstly, Leslie's scenario assumes continuous, rapid growth (i.e., where each successive sub-population is ten times larger than the last). We can reject this assumption. The Devil's Shooting Room is disanalagous insofar as the real world is peppered by a catalogue of astrophysical processes which severely restrict how often a population explosion can take place, as well as how much of a spike in the population's size we can expect. After all, there are limits to how many persons can exist on Earth, and even if we spread out across the galaxies (occupying multiple cosmological horizons in the long-run), there are limits to how much energy is found in these horizons, especially the further down the line we look (e.g., the Dark Era).[13] Secondly, it's not true of our world that the risk of being 'shot' is uniform across time. There will be epochs when the risk of extinction is astronomically higher than, for example, during the cave-man stage of our cosmic evolution. Similarly, we might find that the risk of going extinct by mankind's own hand is much lower, even negligible, once our population becomes very advanced, having mastered control of our cosmic environment. Of the two, the second objection is more forceful, I think. Indeed, although this kind of problematic reasoning isn't quite a storm in a teacup,[14] it's far from obvious that my interlocutor ought to reason as if he will find himself face to face with DEATH upon leaving the rabbit hole.

[12](Leslie, 1996a, 251)

[13]For example, "[when] the cosmic age exceeds 10^{100} years, the black holes will be gone and the cosmos will be filled with the leftover waste products from previous eras: neutrinos, electors, positrons, dark matter particles, and photons of incredible wavelength. In this cold and distant Dark Era, physical activity in the universe slows down, almost (but not quite) to a standstill. The available energy is limited and the expanses of time are staggering, but the universe doggedly continues to operate. Chance encounters between electrons and positrons can forge positroniums atoms, which are exceedingly rare in an accelerating universe. In addition, such atoms are unstable and eventually decay. Other low-level annihilation events also take place, for example, between any surviving dark matter particles. In the poverty of this distant epoch, the generation of energy and entropy becomes increasingly difficult (Adams, 2008, 41-42)."

[14]The Doomsday Argument gets generated from similar-type (anthropic) reasoning. For a good introduction see (Leslie, 1996a, 187ff); (Bostrom, 2002, 89ff). For the No-Outsider Requirement Solution,

see (Bostrom, 2002, 112-115); (cf. Ćirković and Milošević-Zdjelar, 2003a). A similar problem similar arises for those that start from different priors (having instead adopted the self-indication assumption) and argue they've avoided the Doomsday Argument's conclusion—see (Grace, 2010). For a quantum flavour of the problem, turn to (Leslie, 1996b).

Appendix D

Life After Extinction

> Piglet sidled up to Pooh from behind.
> "Pooh!" he whispered.
> "Yes, Piglet?"
> "Nothing," said Piglet, taking Pooh's paw. "I just wanted to be sure of you.
>
> A. A. Milne, *The House at Pooh Corner*

In subsection 4.4.1.6, *The Folly of Anti-Natalism*, I said we were going to ignore those outcomes in which humanity goes intentionally extinct by killing itself. The function of appendix D is to double-down on this claim. In a nutshell, even if the act of exterminating humanity were feasible, life might find a way to rise from the ashes. If so, then we should be extra cautious when executing an anti-natalist plan. Life might restart under even worse conditions than were faced by our cave-dwelling ancestors.[1] This depends in large on how we brought about our own doom. If we released a terrible pathogen which polluted the world, then this new sub-population might find itself in a broken world from which there is no escape.

I think life could be far worse post-doomsday if an anti-natalist plan is executed. But this isn't the claim that I will be defending in this short notice. My goal is to convince you that it isn't all that hard for intelligent life to get started in this post-apocalyptic world. Let's call this the *Galactic-Goldilocks Hypothesis*. Conditional on this claim's truth, we *might* want to flesh out our policies so that steps are taken to improve the starting conditions of this possible sub-population.[2] And we could test this out by supplementing my toy model with (a) a hostility parameter (described

[1] Bear in mind, because his predecessors will have affected both his world and place in cosmic history, we take them to belong to a single population.

[2] This is different from the scenario where (e.g.) the Earth is dying, and the last of us may, in a gambit to save the future of humanity, send off generation-pods—i.e., pods which contain the seeds of humanity, as well as an artificial intelligence designed to 'plant & tend them'—in every direction where there may be habitable planets. Someone may think it's plain that we ought to send the generation-pods, but disagree that we ought to take costly steps during the end days to preserve the preconditions required for life's origin and evolution, or the habitability of our planet, solar system, galaxy, and so on.

below); and (b) a coefficient which describes how many 𝔼s were lost in the poverty of the interim (between our doomsday and life re-emerging).[3] If these features were built into my toy model, then it seems likely that the anti-natalist plan would lose terribly (in terms of maximizing expected fairness) to the alternatives. But I have not run the toy model this way, so cannot be sure.

*
* *

Let's begin by considering the resilience of the microbial world here on Earth. Many—too many—species have gone extinct on this planet after several revolutions around the Milky Way. Indeed, our world has been rocked by several mass-extinction events. But microbes (single cell creatures) have survived. And there is microbial life that has been found thriving under conditions previously considered impossible.[4] It seems nothing short of the world ending as the Sun turns into a red giant in about five billion years can guarantee their extinction. For example, even a supernovae explosion that takes place within 30 light years won't do the trick, given that microbial life persists in the deep subsurface of our planet. Furthermore, we still aren't sure if microbial life will be found on barren planets like Mars.[5]

This means that the basic building blocks required for intelligent life's evolution are preserved. Even if a mass-extinction event were to occur (as it has several times before), the world will recover (as it has before),[6] and microbes will continue populating it. The pressing consideration now is whether the evolutionary steps in-between single-cell life and complex, intelligent life are too difficult (or idiosyncratic) to be successfully navigated again during the Earth's (relatively) short existence in the cosmos. There are two rudimentary hypotheses I'll consider here, neither of which gets things 'just right', unlike the Galactic-Goldilocks Hypothesis. These are the *Rare Earth Hypothesis & Extended Continuity Thesis*. I start with the *Rare Earth Hypothesis*.

[3]Some long-run 𝔼s will undoubtedly remain. It's not as if, for example, the eradication of malaria stops being the case just because humanity went extinct. However, other 𝔼s will, of course, slowly or instantly deteriorate; and if so, the new sub-population must re-invent them (e.g., gene-editing). Obviously, we shouldn't reset the clock to zero just because the prior sub-population went extinct. But how well-off the population that emerges from the ashes finds itself is nebulous. I suppose this largely depends on how devastating the extinction event that wiped out the prior population proved to be, as well as what if any safety-engineering the earlier sub-populations undertook. Moreover, my guess is that the longer it takes for the new sub-population to rise from the dead so to speak, the more of these long-run goods they will not have immediate access to. The new sub-population may have to re-do very many of the steps their forebears had already done. Even if malaria is long gone, buried in the past, (*inter alia*) global hunger may once again mar this civilization.

[4]See especially (Cockell, 2003).

[5]Cockell discusses this topic at length in (Cockell, 2003, chp. 9).

[6]Importantly, even if the world is irreversibly *broken*—e.g., a toxic bog—in some way, we must concede that things are indeed very strange in our universe. It is, at the very least, nomologically possible for life to take highly weird, unexpected forms, especially given our poor (current) understanding of 'life' and its limits.

D.1 Rare Earth Hypothesis

According to the Rare Earth Hypothesis, "while simple microbial life is probably ubiquitous throughout the galaxy, complex biospheres, like the terrestrial one, are very rare due to the exceptional combination of many distinct requirements".[7] For example, it's believed that (a) rare nuclides need to be present in the planetary interior in sufficient amounts to enable plate tectonics and the functioning of the carbon-silicate cycle; (b) the circumstellar habitable zone must be established; and several more (while recognizing that the list is not yet complete).

However, the general reasoning boils down to two points. First, these conditions are considered mostly independent and a priori unlikely, so that their combination is very rare and probably unique in, at least, the Milky Way.[8] Second, the Rare Earth Hypothesis accepts that the *astrobiological landscape*—an abstract landscape-like structure in the space of astrobiological parameters capable of unifying a set of different, viable evolutionary histories of life in any particular region of space[9,10]—is such that there is no big gap between non-living matter and life. That is to say, the evolution of complex (intelligent) life is cosmologically rare, but biogenesis (of simple, microscopic life forms everywhere) can happen rather quickly where physical, chemical, geological etc. conditions are satisfied.[11] So, if we found traces of microbial life on Mars, for example, or biomarkers on another extrasolar planet, then this would offer no evidence that complex life isn't rare. Rather, the Rare Earth Hypothesis would be falsified if we were to find simple life forms absent in the galaxy (or, alternatively, by finding traces of complex life forms elsewhere).

As Milan Ćirković points out in The Astrobiological Landscape, there are essentially three lines of attack regarding the Rare Earth Hypothesis. We could (a) deny the independence of the various requirements; (b) deny that the particular requirements are unlikely to be met; or (c) argue that the terrestrial biosphere might not be representative of the entire class of biospheres capable of giving life to intelligent creatures (such that

[7] (Ćirković, 2012, 131)

[8] (Ćirković, 2012, 131-132)

[9] (Ćirković, 2012, 78-79)

[10] As covered in (Ćirković, 2012, 80-81), different evolutionary histories of life include: (a) *Dead Space* (i.e., the galaxy is entirely dead and stays that way in all epochs); (b) *Sporadic Life* (i.e., life emerges here and there, without any particular correlation either in space or time); (c) *Rare Earth*; (d) *False Precision* (i.e., the number of inhabited planets in the galaxy behaves like $A \ln(\frac{t}{t_{MW}}) + B$, where t_{MW} is the present age of the Milky Way, and A, B are given constants); (e) *Red Dwarf Kingdom* (i.e., while simple life forms are ubiquitous in the galaxy, peaks of complexity could be found only around M-class red dwarf stars, which create the most stable conditions over their huge lifetimes; (f) *Galactic Club* (i.e., life, intelligence and civilizations evolved independently long ago in various places, and many older civilizations are aware of one anotehr and are actively communicating and/or collaborating); (g) *Extinct Galactic Club* (i.e., the same as in *Galactic Club*, except that some unspecified natural or intentioanl cataclysm has destroyed most or all advanced civilizations at some moment in the past (rather close to the present in comparison to t_{MW})); and (h) *Black Clouds with a Vengeance* (i.e., life on planets is a rare exception and most astrobiological complexity lies within giant molecular clouds and their low-density, low-temperature ecosystems).

[11] (Ćirković, 2012, 132)

the terms in the Rare Earth Equation would be considered too restrictive). In his book, Ćirković makes a compelling case for (a), which I don't propose to rehearse here.[12] Myself, I find that (c) is strong enough. Indeed, we have already uncovered scientific models that undermine, sometimes completely, some of the conditions listed by proponents of the Rare Earth Hypothesis. For example, having a giant planet ('Jupiter') at the right distance to deflect much of the incoming cometary and asteroidal material is taken to be a strict requirement. However, the work of Horner and Jones shows that this is untrue in a large part of parameter space.[13] Moreover, they conclude "that such planets often actually increase the impact flux greatly over that which would be expected were a giant planet not present".[14] So, Jupiter might actually have a detrimental effect on the habitability of our green home planet! It seems awfully early—and hence, imprudent—for us to start ruling out just how weird things might be out there—especially if we begin rejecting (nomologically) possible exotic life forms from a meagre sample size of one.

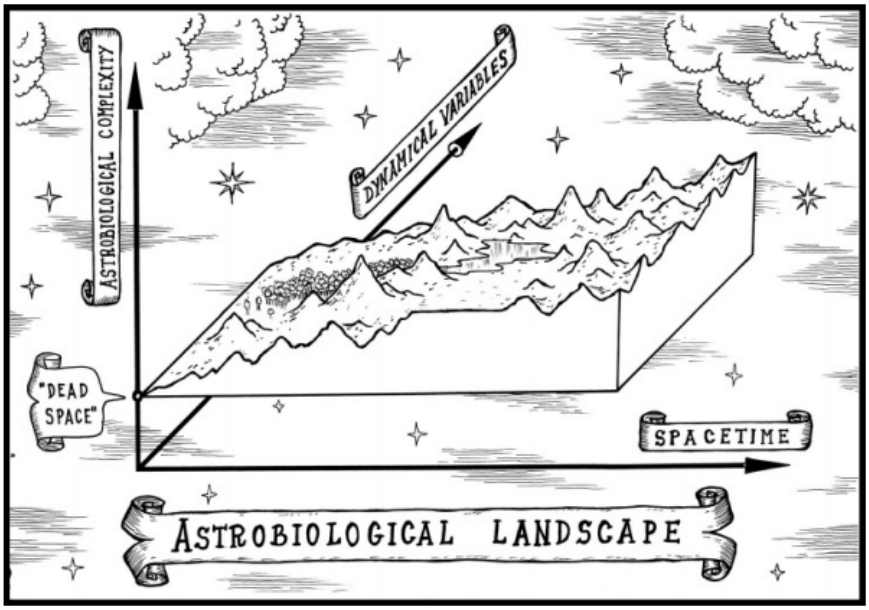


Figure D.1: Astrobiological Landscape - While this does not capture the multidimensional parameter space that truly undergirds the astrobiological landscape, here we can see spacetime as one parameter, and all (so far poorly understood) astrobiological dynamical parameters condensed into another. The model encompasses all the different variables of relevance for habitability (e.g., the temperature range for finding liquid water), while appreciating that these variables should have some give, bend; producing a diverse range of astrobiological complexity in our universe. (Source: (Ćirković, 2012, 80))

But we might further be sceptical with respect to particular requirements being

[12] Turn to (Ćirković, 2012, 133-138) for his "unphysical *ceteris paribus*" objection.
[13] (Horner and Jones, 2008).
[14] (Horner and Jones, 2009, 75).

a priori unlikely, ours being the one world where this rare phenomena succesfully occurred. Take, for example, the following instructive passage on Levinthal's paradox.[15]

> How long does it take for a protein to fold up into its native structure? In a standard illustration of the Levinthal paradox, each bond connecting amino acids can have several (e.g., three) possible states, so that a protein of, say, 101 amino acids could exist in $3^{100} = 5 \cdot 10^{47}$ configurations. Even if the protein is able to sample new configurations at the rate of 10^{13} per second, or $3 \cdot 10^{20}$ per year, it will take 10^{27} years to try them all. Levinthal concluded that random searches are not an effective way of finding the correct state of a folded protein. Nevertheless, proteins do fold, and in a time scale of seconds or less. This is the paradox.[16]

The amount of time required for the protein-folding phenomena to 'get lucky' is astronomical, making the origin of life practically impossible. Indeed, the random sampling of every possible conformation would take longer than the age of the universe! Yet, here we are.

The paradox disappears once we reject the pathway idea inherent to the gedanken put forth by Levinthal: folding can funnel to a single stable state by multiple routes in conformational space. In short, folding occurs into a single stable state because of parallel—not sequential, as his thought experiment proposed—microscopic multi-pathway diffusion-like processes, such as the formation of α helices. The take-home message being that when a mathematical calculation shows that some routine process is impossible, then it's the calculation that's misguided, or the assumption behind the calculation. One's anthropic stance ought not show that life is rare by depending on its being impossible; intelligent life isn't a miracle. In general, then, we have pro tanto reason to be wary of the Rare Earth Hypothesis' core claim. We cannot exclude the possibility of widespread *evolutionary funneling events*. After all, we do find that there's a small and physically reasonable energy bias against locally unfavorable configurations which reduces the time required for finding the correct state of a folded protein to a biologically significant size.[17] Importantly, there might be, for all we know, very many such evolutionary mechanisms to be found (in at least some astrobiological domains). The evolution of simple microbial life might be getting kicked into action from several such sources.

We must bear in mind here the pitfalls of anthropic bias, of course. We cannot observe conditions that are incompatible with our making those observations. So, we learn nothing about the probability of life's formation—or any one critical step in the formation of life being met—from the observation that we happen to exist in this world.[18]

[15] Eric Drexler provided me with this toy example in private communication.

[16] (R. Zwanzig and Bagchi, 1992, 20)

[17] See esp. (Dill and Chan, 1997).

[18] See especially (Bostrom, 2002). A related worry here is *anthropic shadow*. We systematically underestimate the frequency of catastrophes that destroy or are otherwise incompatible with the existence

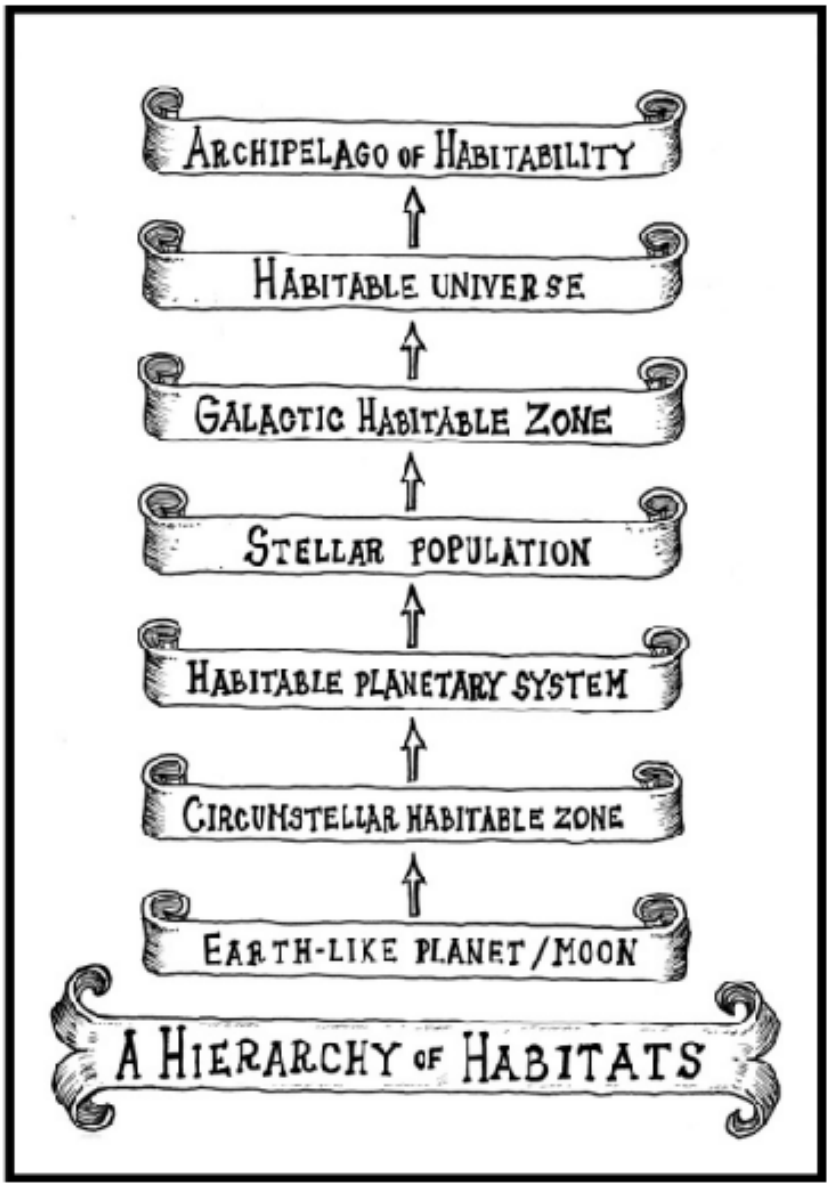


Figure D.2: A Heirarchy of Habitats - From smallest to largest, here one will find the order of habitats where life can take place. Clearly, the lower habitats depend on the higher habitats—e.g., one won't find a stellar population in an uninhabitable universe. (Source: (Ćirković, 2012, 97).)

My response is to grant the process of evolutionary funneling might itself be very unlikely from a randomly selected sample of all possible universes. But the general lesson that *can* be taken away from the toy example is that the successful emergence of intelligent creatures on Earth is more easily explainable if we were pushed through these critical steps by—and their alignment having been sealed by—certain hierarchical processes; in other words, that this success is not as improbable *on this planet* as previously imagined. Again, one's anthropic stance ought not show that life is rare by depending on its being impossible. More so, processes such as evolutionary funneling aren't, at first blush, fixed on any particular astrobiological domain or cosmic epoch. Proteins on an exo-planet in an alien galaxy will also abide by an energy bias. So, we can say, in defense of (b), that some credence ought to be preserved in those mechanisms or processes that will continue urging life along its way on Earth (the Milky Way, and perhaps even higher up the *Hierarchy of Habitats*) even after our world has been rocked by a catastrophe of biblibcal proportion. It's premature to suppose otherwise.

To sum up, the terrestrial biosphere might not be representative of the entire class of biospheres capable of giving life to intelligent creatures, *and* we have solid reason to doubt that the requirements for intelligent life to take root are unlikely to be met. All in all, the Rare Earth Hypothesis looks to be bunk. However, it would be equally imprudent for us to overreact, throwing out the baby with the bathwater. Specifically, we shouldn't reject the idea of dynamical mechanisms that explain the number of observers under the widest possible range of environmental parameters (that is to say, an *astrobiological landscape*). "The truth is, as usual, somewhere in between: neither reification of habitable zones, nor vague preaching about the ubiquity of strange life. We need quantitative models, out of which habitable zones would emerge as just rough, rule-of-thumb approximations."[19]

Let's now turn to the second hypothesis about the different pathways concerning life's evolution in our world. According to Ćirković's *Extended Continuity Thesis*, "there are no unbridgeable gaps between simple life and a complex one, and between complex life and an intelligent one (possibly between biological intelligent life and a postbiological intelligent one). Whenever and wherever physical, chemical, ecological, etc. conditions are suitable, the emergence of complex life is highly probable—and ditto for intelligent life."[20,21] If this were right, then we ought to also accept the feasibility

---

of observers. Essentially, our overconfidence grows to be very large for very destructive threats. See (Ćirković et al., 2010).

[19](Ćirković, 2012, 139)

[20](Ćirković, 2012, 143)

[21]He offers a defense of the extension, which we won't explore here, at (Ćirković, 2012, 143-147). He imputes the bare-bones skeleton of his thesis to Iris Fry's *Continuity Thesis*. According to Fry, there is no unbridgable gap between inorganic matter and living systems, and that *under suitable physical conditions the emergence of life is highly probable* (Fry, 1995, 389). Important to bear in mind is that her thesis doesn't amount to a recasting of Barrow and Tipler's point (that the universe must have those properties which allow life to develop within it at some stage in its history). Her thesis is a description of the evolutionary proess, not resting on any cosmological model over another—indeed, it's compatible with an empty universe in which there is no physical condition for evolution to start, and so no life or observers to be found (Ćirković, 2012, 142). Finally, we should bear in mind that

of the aforemetioned scenario. It's possible for intelligent life on Earth to get wiped out (e.g., by nuclear winter), the world to heal,[22] and another intelligent form of life to evolve from the only survivors of the catastrophe, microbial life. More so, it is possible for a population to have colonized some chunk of the cosmological horizon before having gone extinct. If the conditions required for life to take off were either already in place or if terraforming instead were performed, then it's possible for these colonists to get snuffed out, their new home to heal, and intelligent life to once again rise from the ashes.

D.2 Lots of Earths: *Rare in Time, Not Space*

But one might wonder, upon gazing at the stars, 'well, where is everyone?'—how might we explain away the Great Silence?[23] If the formation of life out of inorganic matter is so probable, then how come we seem to be all alone? Clearly, something must be said about the sheer lack of evidence of other intelligent creatures existing in our light cone if we wish to preserve the plausibility of the hypothesis.[24] To this end, I propose colouring in the Extended Continuity Thesis by adding (a) the *Too Early Objection* from the same author, Milan Ćirković, and (b) its corollary, the *Problem of Too Late*; culminating in the Galactic-Goldilocks Hypothesis.

In a seminal paper, Ken Olum provides us with several possible solutions to where Fermi's paradox goes wrong.[25] Among them are:

there may be very many 'critical steps' in getting from inorganic matter to microbial life or colonizing the galaxies. Some of these might take longer than others to pass through. (We will count getting past the age of (man-made) existential risk as one such critical step.) Even so, the formation of life (and its evolution) is highly probable, according to our thesis, so long as conditions are suitable (pace the Rare Earth Hypothesis).

[22] Actually, to hold this position we must also add the following qualification: there is sufficient time left before the Sun turns into a red giant for intelligent life to evolve from microbial life. Not knowing how long the evolutionary process might take (again, given our sample size of one), a (somewhat) shorter time-frame than our own history suggests ought to be acceptable. Moreover, from behind the veil of ignorance, very many different timelines will be considered by interlocutors; some of which will include a shorter stewing period before evolution really took off (as it did during the Cambrian explosion).

[23] (Brin, 1983)

[24] There might be such intelligent creatures outside of our past light cone. See (Wesson, 1990).

[25] To be sure, he is offering possible solutions to a slightly different problem (i.e., that anthropic reasoning (of a sort) suggests the overwhelming probabilistic prediction that we live in a super-civilization—which fails spectacularly upon observing the world around us). See (Olum, 2004). However, the proposed list is instructive even if it's far from being exhaustive. For example, some have discussed the 'deadly probes scenario' (Brin, 1983); (cf. Sandberg and Armstrong, 2013a) (see also (Sandberg and Armstrong, 2013b), as well as (Sandberg, forthcoming) for a model optimized for speed). A more recent explanation which didn't make Olum's list is that it may be rational to aestivate—note that aestivation is the proper term for hibernation performed through warmer periods—until a later epoch. "Even if only the resources available in a galactic supercluster are exploited, later-era exploitation produces a payoff far greater than any attempt to colonize the rest of the accessible universe and use the resources early. In fact, the mass energy of just the Earth itself ($5.9 \cdot 10^{24}$ kg) would be more than enough to power more computations than could currently be done by burning the present observable universe!"

1. The universe is not infinitely large.
2. Colonization of the galaxy is not possible.
3. We are a 'lost colony' of a large civilization.
4. And if none of these is very likely on its own, then some combination might alleviate the problem.

In a section titled 'Infinitesimally few civilizations become large', Olum writes, "[something] must be wrong with our understanding of how civilizations evolve if only one in a billion can survive to colonize its galaxy".[26] But the correct title, as Ćirković identifies, should have been 'Infinitesimally few civilizations have become large *so far*'.[27] Milan goes on to argue that

> [the] devil hides in the details. ... There is no inconsistency here. The universe, be it infinite or finite, evolves: it changes with cosmic time. What has been a sufficient condition for X at epoch t_1, need not be sufficient at epoch t_2.[28]

More so,

> [we] expect intelligent observers to arise only within a well-defined temporal window of opportunity. Since its appearance against the background cosmological time is an observation selection effect, we shall dubb it the anthropic window.[29,30] Indeed, there are strong empirical grounds on which we can safely claim that the universe has been less hospitable to life in earlier times.[31]

(Sandberg et al., forthcoming, 9) However, Sandberg, Armstrong, and Ćirković go on to tease out a few limitations on hibernation. Primarily, the success of hibernating presupposes that the civilization has solved urgent survival issues (Sandberg et al., forthcoming, 3). A sub-population that has yet to push down anthropogenic risk below some threshold has solid basis for postponing hibernation insofar as it should discount the far future due to uncertainty. More so, hibernating too early is foolish. If astrophysical processes are such that the hostility parameter introduced earlier, τ, is still very high, then they ought to wait—or take steps to wrap themselves in cotton wool (e.g. an advanced artificial intelligence system that observes cosmic conditions, and, if needed, either moves the population to another local region or wakes them up—before stepping into their hibernating chambers). A well-executed hibernation furthermore presupposes that the population can retain control over its volume against alien hypercivilizations.

[26](Olum, 2004, 5)

[27](Ćirković, 2006, 373) (my emphasis)

[28](Ćirković, 2006, 373-374)

[29](Ćirković, 2004b, 55)

[30]The boundaries of the anthropic window, of course, aren't well understood, given (as noted earlier) our poor grasp of the physical, chemical, and biological pre-conditions of observership (Ćirković, 2004b, 55).

[31]cf. (Lineweaver, 2001)

Take, for example, chemical enrichment in that early period; "fewer elements heavier than helium meant a smaller probability for the formation of terrestrial planets, and perhaps a smaller probability of biochemical processes leading to life, intelligence and observers."[32] Furthermore, gamma-ray bursts, collosal explosions caused either by terminal collapse of supermassive objects or mergers of binary neutron stars, and other cosmic collisions, have acted to thwart or impede the (uniform) formation of life in our galaxy.[33]

This can be made more precise by formalizing the phenomena. Let's describe the cumulative effect of all naturally-occuring filters on (or critical steps for) our evolutionary sequence (leading from cave-dwellers to galaxy-colonizers) as a *hostility parameter*. This is, of course, a narrow slice of the full range of threats to intelligent life in the galaxy. After all, some man-made catastrophes have the ability to destroy *all* life (including alien) within our cosmological horizon.[34] However, we are for now only interested in explanations for the absence of intelligent life teeming the observable universe that don't rely on hostile alien lifeforms or ancient (Earth-originating) civilizations which may have released (e.g.) deadly probes throughout the stars *and* whose remains have somehow eluded being uncovered by archaeologists.

The hostility parameter has its spatial and temporal distributions, and these act as constraints on any a-posteriori estimates of the probability of an interlocutor finding himself in a large or small population upon leaving the rabbit hole. Following Ćirković, let's imagine we will find ourselves in a habitable region of the *Archipelago of Habitability* (i.e., the set of regions in parameter space describing those parts of the multiverse that are hospitable to life and intelligent observers of any kind).[35,36] Taking a representative volume of space and having counted N_0 habitable sites where life and intelligence can develop, let's assume the probability of a small civilization becoming a large one evolves over time as:

$$p_l(t) = p_{l0}[1 - \exp(\tfrac{-t}{\tau})], \tag{D.1}$$

where t is the cosmic time from as early as the galaxy's formation (for instance), τ is the hostility parameter, and p_{l0} is the asymptotic 'standard' probability, *ceteris paribus*, of a small civilization making the transition to a large one.[37] Essentially, the cosmos' hostility towards life goes down as we move further downhill from the galaxy-formation epoch. Moreover, D.1 informs us of just how probable the convalescence from a huge

[32](Ćirković, 2012, 74)

[33]See (Ćirković, 2012, 77ff); (Ćirković, 2006); (Ćirković and Vukotić, 2008); (Ćirković et al., 2009).

[34]E.g., if we were to trigger a vacuum-phase transition when conducting a high-energy particle collider experiment.

[35]The sub-section relies heavily on the work produced at (Ćirković, 2012, 75-76, 88-92)

[36]"The Archipelago is a subset of the populated landscape of either string theory or any other overarching 'Theory of Everything' with multiple low-energy solutions ('vacua')" (Ćirković, 2012, 89). See figure D.3.

[37]In the astrobiological case, of course, there is no standard probability based on prior statistics for us to pull from here—that is, no such prior equilibrium exists (apart from, trivially, the dead galaxy before planet formation began) (Ćirković, 2012, 75).

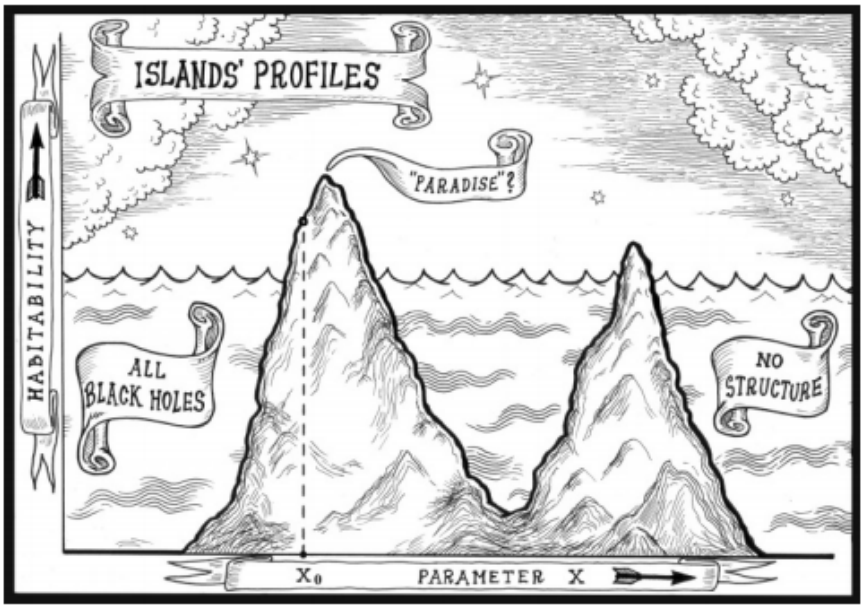


Figure D.3: Archipelago of Habitability: Parameter *X* - Two possible habitable "islands" are shown here, as well as domains where life cannot take hold (for example, if all matter is in the form of black holes or where there had been no structure formation at all). (Source: (Ćirković, 2012, 91).)

adverse perturbation to the norm (p_{l0}) is for a population. If the fraction of observers living in a large civilization is taken to be the following[38]

$$f_{\text{large}}(t) = \frac{N_0 p_l(t)\langle n_l\rangle}{N_0 p_l(t)\langle n_l\rangle + N_0[1-p_l(t)]\langle n_s\rangle}, \tag{D.2}$$

then, by combining the two equations, we obtain

$$f_{\text{large}}(t) \approx \frac{\langle n_l\rangle}{\langle n_l\rangle + \frac{1}{p_{l0}}\cdot\frac{1}{[1-\exp(\frac{-t}{\tau})]}\langle n_s\rangle}. \tag{D.3}$$

Equation D.3 tells us that we are bound to approach an "equilibrium state in which perturbations from past large-scale physical processes (like nucleosynthesis and γ-ray bursts) [cease] to play a significant role, and the only parameter describing the transition between small and large civilizations is the asymptotic probability, p_{l0}.[39] And if so, then gamma-ray burst rarefaction, for instance, also means the chance of finding other large populations in our cosmological horizon gradually increases.

This presents with a far better response than Olum's to the problem. Insofar as we have yet to find traces of alien life, we can conclude, from equation D.3, that *at*

[38] Here, n_l expresses the average number of observers in a large population, and n_s does so for a small population.

[39] (Ćirković, 2012, 75)

least one of the following propositions must be true: (a) all observers find themselves in a large population (i.e., $f_{\text{large}}(t) \approx 1$); (b) the asymptotic probability of evolving into a large population is nill (i.e., $p_{l0} \approx 0$); or (c) the hostility parameter complements cosmic time in such a way that it pulls down $p_l(t)$ to ≈ 0 (i.e., $\exp(\frac{-t}{\tau}) \approx 1$).[40] Clearly, (a) is paradoxical. (b) is also impoverished given that it presupposes some kind of inherent, fated doom and gloom (or refusal to engage in interstellar travel) of practically all populations that might ever exist in the universe. (c) does much better. A *global evolutionary effect* acting to impede the formation of large populations fits our observations, unlike (a), and implies nothing strange about alien sociology, as does (b).[41,42]

This in hand, we finish our response to the Great Silence by turning to James Annis' phase-transition model. In an instructive passage, Annis writes,

> [Since] the regulation mechanism exhibits secular evolution, with the rate of catastrophic events decreasing with time, at some point the astrobiological evolution of the Galaxy will experience a change of regime. When the rate of catastrophic events is high, there is a sort of quasi-equilibrium state between the natural tendency of life to spread, diversify, and complexify, and the rate of destruction and extinctions. When the rate becomes lower than some threshold value, intelligent and space-faring species can arise in the interval between the two extinctions and make themselves immune (presumably through technological means) to further extinctions, and spread among the stars. Thus the Galaxy experiences a phase transition: from an essentially dead place, with pockets of low-complexity life restricted to planetary surfaces, it will, on a very short (Fermi-Hart-Tipler) timescale, become filled with high-complexity life.[43]

So how are we the only ones so far? Well, if Ćirković and Annis are right, then we are living within that interval, on the verge of a galactic phase transition. In other words, it's *too early* for there to be explosive life spreading out among the stars.

A corollary of the *too early* objection is that there will be an epoch after which it's too late for life to evolve from inorganic matter. Eventually the grist for getting life out of the inorganic-windmill will become sparser as dark energy becomes the dominant

[40](Ćirković, 2012, 76)

[41](Ćirković, 2012, 76)

[42]The toy model can be complicated by introducing several 'critical steps' (or filters) in evolving from inorganic matter all the way through to a galaxy-colonizing super-population. To do this we would substitute a single term in the denominator of D.3 with something like the following ((Ćirković, 2012, 76)):

$$\prod_{i=1}^{n} \frac{1}{1-\exp(\frac{-t}{\tau_i})} \tag{D.4}$$

The Great Silence is, of course, no problem whatsoever if the proposition $p : (\exists i)\tau_i \gg t$ is true (Ćirković, 2012, 76). After all, the probability of a population surviving it's infancy, let alone coming into being would be very, very low. But p won't be true of all cosmological epochs yet ahead.

[43](Ćirković, 2004b, 54); (Annis, 1999)

source of energy in the galaxy. Call this the Problem of Too Late. This conforms with Annis' phase-transition model. There may be very many anthropic windows (such as is proposed by the Galactic Punctuated Equilibrium)[44], but their number is finite.[45] Some later era(s) in physical eschatology will forbid the necessary conditions for life's evolution to align or be met. (Bear in mind, however, that if the population is sufficiently mature to manage their survival in an onslaught of catastrophic risks (both from nature as well as man-made), they can, in principle, exist at an epoch outside the anthropic window.)[46] Plus Annis' phase-transition model is compatible with equation D.3 for the reason that there we are only measuring for the probability of an *existing* small population becoming a large one over time.

*
* *

Together, these two propositions—the Extended Continuity Thesis and Annis' phase-transition model—culminate in the *Galactic-Goldilocks Hypothesis*. It holds that (a) a single population may be broken down into several iterations, separated by great expanses of time (and space), of sub-populations; (b) extinction is permanent if approaching the end of (or if after) the *final* anthropic window closing; and (c) though the hostility parameter might give way to the asymptotic probability, p_{i0}, we ought to be wary of the anthropogenic threats ahead that fall outside this equation. After all, neither p_{l0} nor the hostility parameter, τ, account for the presence of such man-made catastrophes.

If the Galactic-Goldilocks Hypothesis were true, then the anti-natalist plan is *ex ante* even worse of a dark gamble. And this, in turn, strengthens my case for dropping this kind of gruesome plan from consideration.

[44]Roughly, the proposal that there are periods in which astrophysical processes permit life to emerge, as well as remain stable for some time ('stasis'), but that peppered in between are irregular catastrophic events taking place in the galaxy which cause existing populations to go extinct. See (Ćirković et al., 2009).

[45]It makes no difference to my subsequent arguments if we asssume that the final anthropic window closes either well or at the moment of heat death.

[46](Ćirković, 2004b, 56)

Appendix E

A Different Toy Model

> "Well," said Pooh, "what I like best," and then he had to stop and think. Because although Eating Honey was a very good thing to do, there was a moment just before you began to eat it which was better than when you were, but he didn't know what it was called.
>
> A. A. Milne, *Winnie-the-Pooh*

I mentioned at the outset of chapter 4 that there is a rather curious feature of ranking lotteries by *expected total utility divided by expected population size*.[1] Essentially, the size of a population within an outcome (within an uncertain lottery) can in some sense swamp. This is because VA operates on something that strongly resembles the Self-Indicating Assumption. Roughly,

> Given the fact that you exist, you should (other things equal) favor hypotheses according to which many observers exist over hypotheses on which few observers exist.[2]

This has also been called the Thirder Position with reference to the Sleeping Beauty Puzzle.[3]

The purpose of Appendix E is to show what follows if we instead accepted the Halfer Position—basically, let's see what pops out if we crunch the numbers such that the value of lotteries are determined by average expected utility. Of course, this is inconsistent with VA. I discussed how so in chapter 4. But there's no harm in at least peeking under the hood and seeing whether this kitten purrs. Here are the relevant algorithms that I plugged into the toy model.

```
1 import random
2 import math
3
```

[1] (Thomas, 2016, 150)
[2] (Ćirković, 2004c, 3); (Bostrom, 2002)
[3] (Elga, 2000); (Lewis, 2001)

E. A DIFFERENT TOY MODEL

```
n = 10000

PV_fast = None

T = 1.0

def f_fast(B, x, T):

    return (B + (4**4 * (math.log(x)) * T))


def run_model(B, L, T):

    print('Running:        ', 'f_fast (absurd takeoff)')

    aggregate_average = 0.0


    for y in range(1, 1000000000):

        average = 0.0

        B = 100.0

        running_total = 0.0

        L = 1000.0

        D = random.randint(1, 4)

        print('Doomsday occurs at ', L/D)


        for x in range(1, n):

                freak_1 = random.randint(0, 500)

                survival_odds = random.randint(0, 4)

                #print('Freak Early: ', freak_1)

                #print('Survival odds: ', survival_odds)


                running_total = running_total + f_fast(B, x, T)

                #print('Generation: ', x)

                #print('Lifetime welfare level: ', f_fast(B, x, T))

```

```
            #print('          Running total: ', running_total)

            #print('Turbulence: ', T)


            if (x > (L/D)):

                  average = (running_total + (f_fast(B, x, T) * (L - x))) / L

                  print('          Breaking out of model with running_total / x =', average)

                  break


            else:

                  if (freak_1 > 499):

                        if (survival_odds < 1):

                              average = running_total / x if x != 0 else 0

                              print('          Breaking out of model with running_total / x =', average)

                              break

                        else:

                              T = 0.0

                              continue

                  else:

                        T = 1.0

                        continue


      aggregate_average = aggregate_average + average

final_average = aggregate_average / y if y != 0 else 0

print('                  Final score: ', final_average)


```

```
run_model(f_fast , L=1000, T=T)
```

Listing E.1: Toy Model II: *Going Dangerously Fast*

```
import random
import math

n = 10000

PV_slow = None

T = 0.5

def f_slow(B, x, T):

    return ((B + (4**4 * (math.log(x)))) * T)


def run_model(B, L, T):

    print('Running:       ', 'f_slow (absurd takeoff)')

    aggregate_average = 0.0


    for y in range(1, 1000000000):

        average = 0.0

        B = 100.0

        running_total = 0.0

        L = 1000.0


        for x in range(1, n):

                freak_1 = random.randint(0, 500)

                freak_2 = random.randint(0, 2000)

                freak_3 = random.randint(0, 6000)

                survival_odds = random.randint(0, 4)

                severity_of_doomsday = random.random()

                #print('Freak Early: ', freak_1)

                #print('Freak Later: ', freak_2)

```

```
            #print('Freak Out: ', freak_3)
            #print('Survival odds: ', survival_odds)


            running_total = running_total + f_slow(B, x, T)
            #print('Generation: ', x)
            #print('Lifetime welfare level: ', f_slow(B, x, T))
            #print('         Running total: ', running_total)
            #print('Turbulence: ', T)


            if (x < L):
                  if (freak_1 > 499):
                        if (survival_odds < 1):
                              average = running_total / x if x != 0 else 0
                              print('         Breaking out of model with
    running_total / x =', average)
                              break
                        else:
                              T = severity_of_doomsday - 0.5
                              continue
                  else:
                        T = 0.5
                        continue

            elif (x < 4000):
                  if (freak_2 > 1999):
                        if (survival_odds < 1):
                              average = running_total /x if x != 0 else 0
                              print('         Breaking out of model with
```

```
    running_total / x =', average)
103
104                                 break
105
106                             else:
107                                 T= severity_of_doomsday -0.5
108
109
110                     else:
111
112                         T = 0.5
113
114                         continue
115
116
117                 else:
118
119                     if (freak_3 > 5999):
120
121                         if (survival_odds < 1):
122
123                             average = running_total / x if x != 0 else 0
124
125                             print('         Breaking out of model with
    running_total / x =', average)
126
127                             break
128
129                         else:
130                             T= severity_of_doomsday -0.5
131
132                     else:
133
134                         T = 0.5
135
136                         continue
137
138
139
140
141        aggregate_average = aggregate_average + average
142
143        #print('                  Final score: ', aggregate_average)
144
145    final_average = aggregate_average / y if y != 0 else 0
146
147    print('                  Final score: ', final_average)
148
149
150
151 run_model(f_slow , L=1000, T=T)
```

Listing E.2: Toy Model II: *Safe-n-Slow*

```
import random
import math

n = 10000

PV_safe = None

T = 0.5

def f_safe(B, x, T):

    return ((B + (4**4 * (math.log(x)))) * T)


def run_model(B, L, T):

    print('Running:        ', 'f_slow_babymaker (absurd takeoff)')

    aggregate_average = 0.0


    for y in range(1, 1000000000):

        average = 0.0

        B = 100.0

        running_total = 0.0

        L = 1000.0


        for x in range(1, n):

                freak_1 = random.randint(0, 500)

                freak_2 = random.randint(0, 2000)

                freak_3 = random.randint(0, 6000)

                survival_odds = random.randint(0, 4)

                severity_of_doomsday = random.random()

                #print('Freak Early: ', freak_1)

                #print('Freak Later: ', freak_2)

                #print('Freak Out: ', freak_3)

                #print('Survival odds: ', survival_odds)
```

E. A DIFFERENT TOY MODEL

```
54
55
56
57              running_total = running_total + f_safe(B, x, T)
58
59              #print('Generation: ', x)
60
61              #print('Lifetime welfare level: ', f_safe(B, x, T))
62
63              #print('         Running total: ', running_total)
64
65              #print('Turbulence: ', T)
66
67
68              if (x < L):
69
70                      if (freak_1 > 499):
71
72                              if (survival_odds < 1):
73
74                                      average = running_total / x if x != 0 else 0
75
76                                      print('         Breaking out of model with
    running_total / x =', average)
77
78                                      break
79
80                              else:
81
82                                      T = severity_of_doomsday -0.5
83
84                                      continue
85
86                      else:
87
88                              T = 0.5
89
90                              continue
91
92              #else:
93
94                      #average = (running_total + (f_safe(B, x, T) * (4000 - x
    ))) / 4000
95
96                      #print('         Breaking out of model with
    running_total / x =', average)
97
98                      #break
99
100
101
102             elif (x < 4000):
103
```

```
104
105                 if (freak_2 > 1999):
106
107                       if (survival_odds < 1):
108
109                             average = running_total /x if x != 0 else 0
110
111                             print('         Breaking out of model with
    running_total / x =', average)
112
113                             break
114
115                       else:
116
117                             T = severity_of_doomsday -0.5
118
119                             continue
120
121                 else:
122
123                       T = 0.5
124
125                       continue
126
127
128           #else:
129
130                 #average = (running_total + (f_safe(B, x, T) * (10000 -
    x))) / 10000
131
132                 #print('         Breaking out of model with
    running_total / x =', average)
133
134                 #break
135
136
137
138           elif (x < 8000):
139
140
141
142                 if (freak_3 > 5999):
143
144                       if (survival_odds < 1):
145
146                             average = running_total / x if x != 0 else 0
147
148                             print('         Breaking out of model with
    running_total / x =', average)
149
150                             break
151
152                       else:
```

```
                                T=severity_of_doomsday  -0.5

                    else:

                          T = 0.5

                          continue


              else:


                    average = (running_total + (f_safe(B, x, T) * (10000 - x
    ))) / 10000

                    print('          Breaking out of model with running_total
     / x =', average)

                    break



       aggregate_average = aggregate_average + average

       #print('                    Final score: ', aggregate_average)

    final_average = aggregate_average / y if y != 0 else 0

    print('                   Final score: ', final_average)


run_model(f_safe , L=1000, T=T)
```

Listing E.3: Toy Model II: *Safety-First*

The remaining three development curves are provided below.

```
def f_safe(B, x, T):

    return ((B + ((4**(3.39) * (math.log(x))) + (x/(5/2))) * T))
```

Listing E.4: Out-the gate II

```
def f_safe(B, x, T):

    return ((B + (x * (1/(4**(3.82))) * (math.sqrt(x)))) * T)
```

Listing E.5: Sluggish Start II

```

def f_slow(B, x, T):

```

```
4     return ((B + (x/2)) * T)
```

Listing E.6: Steady II

E.1 The Results

After running the toy model, I arrived at the following results (averaged over every iteration under the SSA).

			Safety First		
Takeoffs	*Go Fast*	*Go Slow*	Early-Pull	Mid-Pull	Late-Pull
Absurd	1450	527	823	901	1174
Out-the-Gate	797	383	515	775	1284
Slugglish-Start	132	165	97	318	676
Steady	249	235	214	516	974

Table E.1: Results

What becomes patently clear is that safe-n-slow loses to dangerously fast. There is only one development curve at which it outperforms dangerously fast—i.e, Sluggish-Start—and another where it's in at least in the same ballpark—i.e., Steady. However, safety-first wins in every category except the absurd takeoff. And, more importantly, we see that postponing a population explosion is always beneficial (in expectation).[4]

There are two different conclusions one might reach here. Insofar as there is no clear winner, we can *either* average over the four development curves, and if so, then safety-first is the clear winner, *or* we should endorse a conditional principle. *If upon lifting the veil it turns out that the absurd development curve is accurate, then go dangerously fast—but if not, then go safety-first.* The worry might be that our forebears would not have been able to tell which development curve is actually accurate. To some extent this is also true of us. So I am inclined to say that we go the former route—safety-first wins. And safety-first does better the longer we postpone it. So, in a sense, we just end up with safe-n-slow.

[4]To clarify, an early-pull means the 1,000$^{\text{th}}$ generation goes extinct upon acquiring the energy for 4,000 generations. Mid-pull has the 4,000$^{\text{th}}$ generation do this with their energy for 8,000 generations. And late-pull has the 8,000$^{\text{th}}$ generation do this with an energy pool sufficient for 10,000 generations.

References

Fred C. Adams. Long-term astrophysical processes. In N. Bostrom and M. Ćirković, editors, Global Catastrophic Risks, pages 33–47. Oxford University Press, 2008. 80, 183

Arif Ahmed. Push the button. Philosophy of Science, 79(3): 386–395, 2012a. 29

Arif Ahmed. Dicing with death. Analysis, 74(4):587–592, 2014a. 28, 29

Arif Ahmed. Infallibility in the newcomb problem. Erkenntnis, 80(2):1–13, 2014b. 28

Arif Ahmed. Exploiting cyclic preference. Mind, 126(504): 975–1022, 2017. 121

Arif Ahmed and Huw Price. Artnzenius on 'why ain'cha rich?'. Erkenntnis, 77(1):15–30, 2012b. 29, 42, 43

James Annis. An astrophysical explanation for the great silence. Journal of the British Interplanetary Society, 52: 19–22, 1999. 196

Frank Arntzenius. No regrets, or: Edith piaf revamps decision theory. Erkenntnis, 68(2):277–297, 2008. 29, 39, 43

Frank Arntzenius. Utilitarianism, decision theory and eternity. Philosophical Perspectives, 28:31–58, 2014. 182

Gustaf Arrhenius and Włodek Rabinowicz. The value of existence. In I. Hirose and J. Olson, editors, The Oxford Handbook of Value Theory, pages 424–443. Oxford University Press, Oxford, 2015. 106, 113

J. D. Barrow. The Book of Universes. Vintage, London, 2012. 180

J. D. Barrow and F. J. Tipler. The Anthropic Cosmological Principle. Oxford University Press, New York, 1968. 179

Brian Barry. Contract theory and future generations. Unpublished Manuscript. 30

Brian Barry. Rawls on average and total utility: a comment. Philosophical Studies, 31:317–325, 1977. 15

Brian Barry. Circumstances of justice and future generations. In Richard I. Sikora and Brian M. Barry, editors, Obligations to Future Generations, pages 204–248. White Horse Press, 1978. 16, 21

Simon Beard. Fairness and the future: Evaluating extreme technological risks. unpublished manuscript, forthcoming. 3, 4, 61, 66

Nick Beckstead. On the Overwhelming Importance of Shaping the Far Future. PhD thesis, Rutgers, State University of New Jersey, US, 2013a. 166

Nick Beckstead. A proposed adjustment to the astronomical waste argument. LessWrong, May 2013b. Retrieved from http://lesswrong.com/lw/hjb/a_proposed_adjustment_to_the_astronomical_waste/. 2

Nick Beckstead. How much could refuges help us recover from a global catastrophe? Futures, 72:36–44, 2015. 84

Jens Christian Bjerring. On counterpossibles. Philosophical Studies, 2:1–27, 2013. 48

Nick Bostrom. Anthropic Bias: Observation Selection Effects in Science and Philosophy. Studies in Philosophy. Routledge, New York, 2002. 60, 183, 184, 189, 199

Nick Bostrom. Astronomical waste: The opportunity cost of delayed technological development. Utilitas, 15(3):308–314, 2003a. 2, 3, 163

Nick Bostrom. Are you living in a computer simulation? Philosophical Quarterly, 53:243–255, 2003b. 181

Nick Bostrom. Infinite ethics. Analysis and Metaphysics, 10: 9–59, 2011. 181, 182

Nick Bostrom. Existential risk prevention as global priority. Global Policy, 4(1):15–31, 2012a. 2, 18, 74, 163, 167

Nick Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. Mind & Machines, 22:71–85, 2012b. 28

Nick Bostrom. Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Oxford, 2014. 78, 84, 168

Nick Bostrom, Tom Douglas, and Anders Sandberg. The unilateralist's curse and the case for a principle of conformity. Social Epistemology, 30(4):350–371, 2016b. 1

David Brin. The great silence—the controversy concerning extraterrestrial intelligent life. Quarterly Journal of the Royal Astronomical Society, 24:283–309, 1983. 192

John Broome. Fairness. Proceedings of the Aristotelian Society, 91:87–101, 1990-1991. 4

John Broome. Weighing Goods. Blackwell Publishers, Cambridge, Massachusetts, 1991. 78

John Broome. Ethics out of Economics. Cambridge University Press, Cambridge, 1999. 48, 106, 113

John Broome. Weighing Lives. Oxford University Press, Oxford, 2004. 30, 35, 48, 54, 66, 71, 105, 106, 108, 109, 110, 111, 113, 115, 116, 121, 122

John Broome. Should we value population? Journal of Political Philosophy, 13(4):399–413, 2005. 109

John Broome. Reply to rabinowicz. Philosophical Issues, 19 (1):412–417, 2009. 109

Sarah F. Brosnan and Frans B. M. de Waal. Monkeys reject unequal pay. Nature, 425:297–299, 2003. 3

REFERENCES

Campbell Brown. Priority or sufficiency... or both? Economics and Philosophy, (2):199–220, 2005. 64, 131

Campbell Brown. Prioritarianism for variable populations. Philosophical Studies, 143(3):325–361, 2007. 131

Campbell Brown. Is close enough good enough? unpublished manuscript, forthcoming. 96

Krister Bykvist. Violations of normative invariance: some thoughts on shifty oughts. Theoria, 73(2):98–120, 2007a. 111, 173

Krister Bykvist. The benefits of coming into existence. Philosophical Studies, 135(3):335–362, 2007b. 106, 113

James Cain. Infinite utility. Australasian Journal of Philosophy, 73(3):401–404, 1995. 182

Joseph Carens. The case for open borders. The Review of Politics, 49(2):251–273, 1987. 18, 20

E. Carlson. Consequentialism Reconsidered. Kluwer Academic Publisher, Dordrecht, 1995. 111, 172

M. Ćirković. Physical eschatology (resource letter). Americal Journal of Physics, 71:1–11, 2003b. 73

M. Ćirković and N. Bostrom. Cosmological constant and the final anthropic hypothesis. Astrophysics and Space Science, 274:675–687, 2000. 179, 180, 181

M. Ćirković and V. Milošević-Zdjelar. Extraterrestrial intelligence and doomsday: A critical assessment of the no-outsider requirement. Serbian Astronomy Journal, 166: 122–133, 2003a. 184

M. Ćirković and M. Radujkov. On the maximal quantity of processed information in the physical eschatological context. Serbian Astronomy Journal, 163:53–56, 2001. 142, 180

Milan Ćirković. Forecast for the next eon: Applied cosmology and the long-term fate of intelligent beings. Foundations of Physics, 34(2):239–261, 2004a. 17

Milan Ćirković. Earths: Rare in time, not space? Journal of the British Interplanetary Society, 57:53–59, 2004b. 193, 196, 197

Milan Ćirković. Is many likelier than few? a critical assessment of the self-indication assumption. Epistemologia, 27: 265–298, 2004c. 199

Milan Ćirković. Too early? on the apparent conflict of astrobiology and cosmology. Biology and Philosophy, 21(3): 369–379, 2006. 193, 194

Milan Ćirković. The Astrobiological Landscape. Cambridge University Press, Cambridge, 2012. 84, 187, 188, 190, 191, 194, 195, 196

Milan Ćirković and Branislav Vukotić. Astrobiological phase transition: Towards resolution of fermi's paradox. Origins of Life and Evolution of Biospheres, 38(6):535–547, 2008. 194

Milan Ćirković, Branislav Vukotić, and Ivana Dragićević. Galactic punctuated equilibrium: How to undermine carter's anthropic argument in astrobiology. Astrobiology, 9(5):491–581, 2009. 194, 197

Milan Ćirković, Anders Sandberg, and Nick Bostrom. Anthropic shadow: Observation selection effects and human extinction risks. Risk Analysis, 30(10):1495–1506, 2010. 191

Charles Cockell. Impossible Extinction: Natural Catastrophes and the Supremacy of the Microbial World. Cambridge University Press, Cambridge, 2003. 186

Daniel Cohen. Expected moral actualism. unpublished manuscript, forthcoming. 111, 123, 173, 174, 175

Owen Cotton-Barratt and Toby Ord. Existential risk and existential hope: Definitions. Future of Humanity Institute: Technical Report 2015-1, 2015. 78, 83

Roger Crisp. In defense of the priority view: A response to otsuka and voorhoeve. Utilitas, 23(1):105–108, 2011. 3

John Cusbert. Acting on essentially comparative goodness. Thought, 6(1):1–11, 2017. 120

John Cusbert and Hilary Greaves. Comparing existence and non-existence. unpublished manuscript, forthcoming. 106, 113

John Cusbert and Robyn Kath. A consequentialist account of narveson's dictum. unpublished manuscript, forthcoming. 105, 106, 107, 108, 112, 114, 115, 119, 120, 121

Aisha Dasgupta and Sir Partha Dasgupta. Socially embedded preferences, environmental externalities, and reproductive rights. unpublished manuscript, forthcoming-*b*. 76

Sir Partha Dasgupta. Birth and death. unpublished manuscript, forthcoming-*a*. 20, 21, 76

P. C. W. Davies. Cosmic heresy? Nature, 273:336–337, 1978. 180

P. A. Diamond. Cardinal welfare, individualistic ethics, and interpersonal comparison of utility. The Journal of Political Economy, 75(5):765–766, 1967. 5, 65

Claus Dierksmeier. John rawls on the rights of future generations. In Joerg Chet Tremmel, editor, Handbook of Intergenerational Justice, pages 72–85. Edward Elgar Publishing Limited, 2006. 23, 29

Ken Dill and Hue Sun Chan. From levinathal to pathways to funnels. Nature Structural Biology, 4(1):10–19, 1997. 189

Eric Drexler. Radical Abundance: How a Revolution in Nanotechnology Will Change Civilization. Public Affairs, New York, 2013. 22

Ronald Dworkin. Taking Rights Seriously. Bloomsbury Publishing, London, 1977. 14

A. Guth E. Farhi and J. Guven. Is it possible to create a universe in the laboratory by quantum tunneling? Nuclear Physics B, 339(2):417–490, 1990. 180

REFERENCES

Andy Egan. Some counterexamples to causal decision theory. Philosophical Review, 116:93–114, 2007. 29, 39

Adam Elga. Self-locating belief and the sleeping beauty problem. Analysis, 60:143–147, 2000. 60, 199

Adam Elga. Defeating dr. evil with self-locating belief. Philosophy and Phenomenological Research, 69(2):383–396, 2004. 37

M. Fleurbaey and A. Voorhoeve. Eqgalitarianism and the separateness of persons. Utilitas, 24(3):381–398, 2012. 4, 8, 61, 64, 66

M. Fleurbaey and A. Voorhoeve. On the social and personal value of existence. In I. Hirose and A. Reisner, editors, Weighing and Reasoning: Themes from the Work of John Broome, pages 95–109. Oxford University Press, 2015. 106, 113

Johann David Frick. Making People Happy, Not Making Happy People: a defense of the asymmetry intuition in population ethics. PhD thesis, Harvard University, US, 2014. 108

Johann David Frick. On the survival of humanity. Canadian Journal of Philosophy, 47(2-3):344–367, 2017. 66, 110, 117

Iris Fry. Are the different hypotheses on the emergence of life as different as they seem? Biology and Philosophy, 10: 389–417, 1995. 191

J. Garriga, V. F. Mukhanov, K. D. Olum, and A. Vilenkin. Eternal inflation, black holes, and the future of civilizations. Journal of Theoretical Physics, 39(7):1887–1900, 2000. 180

Jerry Gaus. Public reason liberalism. Unpublished manuscript, forthcoming. 15

Allan Gibbard and William Harper. Counterfactuals and two kinds of expected utility. theoretical foundations. In Clifford A. Hooker, James J. Leach, and Edward F. McClennen, editors, Foundations and Applications of Decision Theory, volume 1. The Western Ontario Series in Philosophy of Science (13), Boston, 1978. 39

Caitlin Grace. Anthropic reasoning in the great filter. Master's thesis, Australian National University, 2010. 184

Hilary Greaves. Cluelessness. Proceedings of the Aristotelian Society, 3(116), 2016. 16, 75, 168, 178

Hilary Greaves. Antiprioritarianism. unpublished manuscript, forthcoming-*a*. 64

Hilary Greaves and Toby Ord. Moral uncertainty about population axiology. unpublished manuscript, forthcoming-*c*. 3, 71

Johan E. Gustafsson. A note in defense of ratificationsim. Erkenntnis, 75:147–150, 2011. 29

Johan E. Gustafsson. Population axiology and the possibility of a fourth category of absolute value. unpublished manuscript, forthcoming. 109, 121

John Halstead. The impotence of a value pump. Utilitas, 27 (2):195–216, 2015. 121

Robin Hanson. The great filter—are we almost past it? online manuscript, September 1998. https://mason.gmu.edu/ rhanson/greatfilter.html. 73

Caspar Hare. Voices from another world: must we respect the interests of people who do not, and will never, exist? Ethics, 117(3):498–523, 2007. 172

R.M. Hare. Rawls' theory of justice—i. The Philosophical Quarterly, 23(91):144–155, 1973a. 20, 30

R.M. Hare. Rawls' theory of justice—ii. The Philosophical Quarterly, 23(92):241–252, 1973b. 16, 17, 18, 19, 30

W.L. Harper, R. Stalnaker, and G. Pearce, editors. Ifs. Reidel Publishing, Dordrect, 1981. 36

J. Harsanyi. Cardinal utility in welfare economics and in the theory of risk-taking. Journal of Political Economy, 61(5): 434–435, 1953. 7

J. C. Harsanyi. Can the maximin principle serve as a basis for morality? a critique of john rawls's theory. The American Political Science Review, 69(2):594–606, 1975. 18

Ori J. Herstein. Why 'nonexistent people' do not have zero wellbeing but no wellbeing at all. Journal of Applied Philosophy, 30(2):136–145, 2013. 49

Nils Holtug. On the value of coming into existence. The Journal of Ethics, 5(4):361–384, 2001. 106

Brad Hooker. Rule-consequentialism, incoherence, fairness. Proceedings of the Aristotelian Society, 95:19–35, 1995. 162

Brad Hooker. Ideal Code, Real World: A Rule-Consequentialist Theory of Morality. Oxford University Press, Oxford, 2000. 161, 162, 169

J. Horner and B. Jones. Jupiter friend or foe? i. the asteroids. International Journal of Astrobiology, 7:251–261, 2008. 188

J. Horner and B. Jones. Jupiter friend or foe? ii. the centaurs. International Journal of Astrobiology, 8:75–80, 2009. 188

Thomas Hurka. Value and population size. Ethics, 93:496–507, 1983. 4

Richard Jeffrey. The sure thing principle. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, 2:719–730, 1982. 27

James Joyce. The Foundations of Causal Decision Theory. Cambridge University Press, Cambridge, 1999. 42

James Joyce. Regret and instability in causal decision theory. Synthese, 187(1):123–145, 2012. 39

Patrick Kaczmarek. How much is rule-consequentialism really willing to give up to save the future of humanity? Utilitas, 29(2):239–249, 2017. 75

Patrick Kaczmarek and Michael Plant. Axiological uncertainty, population ethics, and totalism: Reply to greaves & ord. unpublished manuscript. 3

Shelly Kagan. Normative Ethics. Westview Press, Boulder, 1998. 74, 173

REFERENCES

Shelly Kagan. Do i make a difference. Philosophy & Public Affairs, 39:105–141, 2011. 23

Guy Kahane. Our cosmic insignificance. Noûs, 48(4):745–772, 2014. 3, 163

Robyn Kath. Shortfall Utilitarianism: A Theory for Variable Population Decisions. PhD thesis, University of Sydney, 2016. 120

Gregory Kavka. Rawls on average and total utility. Philosophical Studies, 27:237–253, 1975. 15

Ursula LeGuin. The Ones Who Walk Away From Omelas, pages 254–262. Orion Publishing Group, 2015. 72

James Lenman. Consequentialism and cluelessness. Philosophy & Public Affairs, 29(4):342–370, 2000. 74, 178

James Lenman. On becoming extinct. Pacific Philosophical Quarterly, 83(3):253–269, 2002. 182

John Leslie. The End of the World. Routledge, London, 1996a. 28, 29, 44, 183

John Leslie. A difficulty for everett's many-worlds theory. International Studies in the Philosophy of Science, 10(3): 239–246, 1996b. 184

Ben Levinstein and Nate Soares. Cheating death in demascus. unpublished, 2017. 34, 36, 37, 38, 39, 40, 41, 44, 45, 46, 48

David Lewis. Counterfactuals. Harvard University Press, Cambridge, MA, 1973. 27

David Lewis. Prisoners' dilemma is a newcomb problem. Philosophy & Public Affairs, 8(3):235–240, 1979. 27, 29

David Lewis. Why ain'cha rich? Noûs, 15(3):377–380, 1981a. 28, 29

David Lewis. Causal decision theory. Australasian Journal of Philosophy, 59(1):5–30, 1981b. 29, 36

David Lewis. On the Plurality of Worlds. Blackwell Publishing, Oxford, 1981c. 36, 49, 53

David Lewis. Why conditionalize? In David Lewis, editor, Papers in Metaphysics and Epistemology, pages 403–407. Cambridge University Press, 1999. 121

David Lewis. Sleeping beauty: Reply to elga. Analysis, 61 (3):171–176, 2001. 60, 199

Charles Lineweaver. An estimate of the age distribution of terrestrial planets in the universe: Quantifying metallicity as a selection effect. Icarus, 151:307–313, 2001. 193

David McCarthy, Kalle Mikkola, and Teruji Thomas. Utilitarianism with and without expected utility. MPRA Preprint No. 79315, 2016. 5, 8, 58

Michael McDermott. Utility and population. Philosophical Studies, 42(2):163–177, 1982. 122

Jefferson McMahan. Problems of population theory. Ethics, 92(1):96–127, 1981. 67, 68, 131, 164

Jefferson McMahan. Asymmetries in the morality of causing people to exist. In M. Roberts and D. Wasserman, editors, Harming Future Persons: Ethics, Genetics and the Nonidentity Problem, pages 49–68. Springer, Berlin, 2009. 131, 164

Jefferson McMahan. Causing people to exist and saving people's lives. The Journal of Ethics, 17(1-2):5–35, 2013. 7

Christopher Meacham. Person-affecting views and saturating counterpart relations. Philosophical Studies, 158(2): 257–287, 2012. 122

Qizilbash Mozaffar. Incommensurability or vagueness? a comment on rabinowicz and sugden. Proceedings of the Aristotelian Society, 112(3):333–338, 2012. 115

Tim Mulgan. Transcending the infinite utility debate. Australasian Journal of Philosophy, 80(2):164–177, 2002. 182

Tim Mulgan. Future People. Oxford University Press, Oxford, 2006. 22, 23, 24

Tim Mulgan. Ethics for a Broken World: Imagining Philosophy After Catastrophe. Acumen, Durham, 2011. 20, 22, 23, 24, 26

Tim Mulgan. Utilitarianism for a broken world. Utilitas, 27: 92–114, 2015. 20, 162, 164, 169

Jan Narveson. Utilitarianism and new generations. Mind, 76: 62–72, 1967. 71

Jan Narveson. Moral problems of population. Monist, 57(1): 62–86, 1973. 110

Yew-Kwang Ng. Some broader issues of social choice. In P. Pattanaik and M. Salles, editors, Social Choice and Welfare, pages 151–174. North-Holland Publishing Company, 1983. 8

Yew-Kwang Ng. What should we do about future generations? Economics and Philosophy, 5:235–253, 1989. 4

Ken Olum. Conflict between antrhopic reasoning and observation. Analysis, 64:1–8, 2004. 192, 193

Onora O'Neill. Constructivism in rawls and kant. In Samuel Freeman, editor, The Cambridge Companion to Rawls, pages 347–367. Cambridge University Press, Cambridge, 2003. 15

Michael Otsuka. Determinism and the value and fairness of equal chances. Talk Delivered at London School of Economics, Department of Philosophy, Logic and Scientific Method, 2017. 4, 64

Michale Otsuka and Alex Voorhoeve. Why it matters that some are worse off than others: An argument against the priority view. Philosophy & Public Affairs, 37(2):171–199, 2009. 3

Derek Parfit. Reasons and Persons. Oxford University Press, Oxford, 1984. 3, 16, 34, 67, 68, 69, 70, 71, 106, 113, 116, 163, 165

Derek Parfit. Overpopulation and the quality of life. In Peter Singer, editor, Applied Ethics, pages 145–164. Oxford University Press, 1986. 2, 83, 96

Derek Parfit. Equality or priority? The Lindley Lecture, The University of Kansas, November 1991. 64

Derek Parfit. Equality and priority. Ratio, 10(3):202–221, 1997. 64

Derek Parfit. On What Matters Vols. I & II. Oxford University Press, Oxford, 2011. 16, 20, 30, 163

Derek Parfit. Another defence of the priority view. Utilitas, 24(3):399–440, 2012. 64

Derek Parfit. On What Matters Vol. III. Oxford University Press, Oxford, 2017. 83

Derek Parfit. Towards theory x: Parts i & ii. unpublished manuscript, forthcoming. 67

Josh Parsons. Axiological actualism. Australasian Journal of Philosophy, 80(2):135–147, 2002. 111

Judea Pearl. The sure-thing principle. Technical Report, 2016. R-466. 27

Judeal Pearl. Causality. Models, Reasoning, and Inference. Cambridge University Press, New York, 2009. 36

Ingmar Persson. Inclusive Ethics: Extending Beneficence and Egalitarian Justice. Oxford University Press, Oxford, 2017. 95, 96

Michael Pressman. A defense of average utilitarianism. Utilitas, 27(4):389–424, 2015. 67, 69, 71

Duncan Pritchard. Risk. Metaphilosophy, 46(3):436–461, 2015. 81, 82

Theron Pummer. The worseness of nonexistence. In E. Gamlund and C. T. Solberg, editors, Saving Lives from the Badness of Death. Oxford University Press, forthcoming. 106, 113

Jonathan Quong. Liberalism Without Perfectionism. Oxford University Press, Oxford, 2011. 15

A. Szabo R. Zwanzig and B. Bagchi. Levinthal's paradoxx. Proc. Natl. Acad. Sci. USA, 89:20–22, 1992. 189

Włodek Rabinowicz. Broome and the intuition of neutrality. Philosophical Issues, 19(1):389–411, 2009a. 109

Włodek Rabinowicz. I—incommensurability and vagueness. The Aristotelian Society Supplementary Volume, 83(1): 71–94, 2009b. 115

John Rawls. A Theory of Justice. Harvard University Press, Cambridge, Massachusetts, 1971. 14, 15, 16, 17, 18, 19, 21, 22, 23, 30, 76, 108

John Rawls. Justice as fairness: Political not metaphysical. Philosophy & Public Affairs, 14(3):223–251, 1985. 16

John Rawls. Political Liberalism. Columbia University Press, New York, 1993. 14, 16, 22, 23, 24, 29

John Rawls. The idea of public reason revisited. In Samuel Freeman, editor, Collected Papers, pages 573–615. Harvard University Press, Cambridge, Massachusetts, 1999. 15

John Rawls. Justice as Fairness: A Restatement. Harvard University Press, Cambridge, Massachusetts, 2001. 21, 23, 24

Joseph Raz. The Morality of Freedom. Clarendon Press, Oxford, 1986. 110

Melinda Roberts. The asymmetry: A solution. Theoria, 77: 333–367, 2010a. 120, 121, 122

Melinda Roberts. Abortion and the Moral Significance of Merely Possible Persons: Finding Middle Ground in Hard Cases, volume 107 of Philosophy and Medicine Series. Springer, New York, 2010b. 111, 117

Jacob Ross. Rethinking the person-affecting principle. Journal of Moral Philosophy, 12(4):428–461, 2015. 122, 125

Dov Samet. The sure-thing principle in epistemic terms. Technical Report, The Leon Recanati Graduate School of Business Administration, Tel Aviv University, 2015. 27

Anders Sandberg. Space races: Settling the universe *Fast*. unpublished manuscript, forthcoming. 192

Anders Sandberg and Stuart Armstrong. Hunters in the dark: game theory analysis of the deadly probes scenario. poster, 2013a. NAM2013, St. Andrews. 192

Anders Sandberg and Stuart Armstrong. Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the fermi paradox. Acta Astronautica, 89:1–13, 2013b. 192

Anders Sandberg, Stuart Armstrong, and Milan Ćirković. That is not dead which can eternal lie: the aestivation hypothesis. unpublished manuscript, forthcoming. 142, 193

Leonard Savage. The Foundations of Statistics. John Wiley and Sons, Inc., New York, 1954. 27, 35

T. M. Scanlon. What We Owe to Each Other. Harvard University Press, Cambridge, Massachusetts, 1998. 108

David Schroeren. Existential risk reduction—a rawlsian duty of justice to future generations. Unpublished Manuscript, forthcoming. 15, 23

Thomas Schwartz. Rationality and the myth of maximum. Noûs, 62(2):97–117, 1972. 125

George Sher. What makes a lottery fair? Noûs, 14(2):203–216, 1980. 5

Carl Shulman. Population ethics and inaccessible populations. *Reflective Disequilibrium*, August 2014. Retrieved from https://reflectivedisequilibrium.blogspot.co.uk/2014/08/population-ethics-and-inaccessible.html. 71, 87

Theodore Sider. Might theory x be a theory of diminishing marginal value? Analysis, 51:265–271, 1991. 4

Peter Singer. A utilitarian population principle. In M. Bayles, editor, Ethics and Population, pages 81–99. Schenkman Pubishing Co., Cambridge: Massachusetts, 1976. 2

Peter Singer. Practical Ethics (2nd Edition). Cambridge University Press, Cambridge, 1999. 2

REFERENCES

Brian Skyrms. The Dynamics of Rational Deliberation. Cambridge University Press, Cambridge, 1990. 39

Nate Soares and Benja Fallenstein. Towards idealized decision theory. In: arXiv: 1507.01986 [cs.AI], 2015. 34, 36, 38

Robert Stalnaker. A theory of conditionals. In Nicholas Rescher, editor, Studies in Logical Theory, number 2 in American Philosophical Quarterly Monograph Series, pages 98–112. Blackwell Publishing, Oxford, 1968. Reprinted in W.L. Harper, R. Stalnaker, and G. Pearce (Eds.), Ifs, D. Reidel Publishing, Dordrecht: 41–55, 1981. 27

Robert Sugden. Ii—on modelling vagueness—and on *Not* modelling incommensurability. The Aristotelian Society Supplementary Volume, 83(1):95–113, 2009. 115

Cass Sunstein. Worst-Case Scenarios. Harvard University Press, Cambridge, 2007. 21

Larry Temkin. Inequality. Oxford University Press, Oxford, 1993. 62, 107

Larry Temkin. Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning. Oxford University Press, Oxford, 2012. 116, 120, 176

Joaquin Teruji Thomas. Topics in Population Ethics. PhD thesis, University of Oxford, 2016. 8, 14, 15, 49, 50, 51, 57, 58, 59, 60, 65, 66, 112, 118, 125, 127, 199

Phil Torres. Agential risks: A comprehensive introduction. Journal of Evolution & Technology, 26(2):31–47, 2016. 1

Joerg Chet Tremmel. The convention of representatives of all generations under the 'veil of ignorance'. Constellations, 20(3):483–502, 1986. 19, 21

W. Vickrey. Measuring marginal utility by reactions to risk. Econometrica, 13(4):319–333, 1945. 7

Tatjana Visak. An evaluation of meacham's 'theory x'. unpublished manuscript, forthcoming. 124

Alex Voorhoeve and Marc Fleurbaey. Priority or equality for possible people? Ethics, 126:929–954, 2016. 8

David Wasserman. Let them eat chances: Probability and distributive justice. Economics and Philosophy, 12:29–49, 1996. 5, 64

Ralph Wedgwood. Gandalf's solution to the newcomb problem. Synthese, 14:1–33, 2011. 29, 43, 174

Paul Wesson. Cosmology, extraterrestrial intelligence, and a resolution of the fermi-hart paradox. Quarterly Journal of the Royal Astronomical Society, 31:161–170, 1990. 192

Roger White. Evidential symmetry and mushy credence. In T. Szabo and J. Hawthorne, editors, Oxford Studies in Epistemology, pages 161–186. Oxford University Press, 2009. 16

Bernard Williams. Moral Luck: Philosophical Papers 1973-1980. Cambridge University Press, Cambridge, 2012. 15

A.D. Wissner-Gross and C.E. Freer. Causal entropic forces. Physical Review Letters, 10:168702–1–5, 2013. 21

James Woodward. The non-identity problem. Ethics, 96(4): 804–831, 1986. 21

D. Yang. What's wrong with modal conceptions of luck and risk? Erkenntnis, forthcoming. 82

Byeong-Uk Yi. Newcomb's paradox and priest's principle of rational choice. Analysis, 63(3):237–242, 2003. 29

Declaration

I herewith declare that I have produced this dissertation without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. The dissertation has not previously been presented in identical or similar form to any other examination board.

The thesis work was conducted from October 2013 through September 2017 under the primary supervision of Campbell Brown (of the London School of Economics) and Ben Colburn (of the University of Glasgow). Adam Rieger stepped in to cover for Campbell Brown in my final year, and Jappa Pallikkathayil performed as a secondary supervisor during my time at the University of Pittsburgh as a visiting scholar. Also, the Future of Humanity at the University of Oxford provided a most-cozy home on several occasions for myself as a visiting academic during this four year period.

Glasgow,

Patrick Kaczmarek