PERSONS, POPULATIONS, AND VALUE

13 October 2020

Kacper Kowalczyk

St Anne's College, Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity Term 2020

Abstract

This thesis consists of six independent papers on personal identity, population ethics, and value theory. The central theme of this thesis is the unimportance of personal identity. I begin in Chapter 1 by developing new Parfitian arguments for the unimportance of personal identity in morality, paying special attention to their implications in population ethics. In Chapter 2 I analyse Mark Johnston's similar metaphysics-driven challenge to person-based morality, but argue that it is less successful than my own. In Chapter 3 I lay out choices facing egalitarians in population ethics, arguing against some recently prominent forms of egalitarianism. In Chapter 4 I provide new decision-theoretic arguments against deontic constraints on harming, suggesting that it is consequentialism rather than deontology that can better respect persons. In Chapter 5 I introduce transfinite extensions of the familiar value-theoretic principles of transitivity and acyclicity. I use them to try to resolve some key issues in population ethics, concerning the value of creating new people and the procreative asymmetry. In Chapter 6 I aim to support these transfinite principles by analysing their role in the theory of rational choice.

Word count: 60591

Acknowledgements

I thank my supervisors, Ralf Bader and Teru Thomas, for their patience, insight, and support.

I want to thank my fellow graduate students at Oxford and elsewhere, especially Michał Maśny, Tomi Francis, Todd Karhu, Aidan Penn, and Korbinian Rüger.

I am also grateful to Theron Pummer and Johan Gustafsson for helpful feedback and encouragement.

I was financially supported by Wolfson College and the Faculty of Philosophy and then by St Anne's College and the Aristotelian Society.

I finished the first final draft of this thesis in quarantine during a global pandemic. I am grateful to Karolina Wątroba who kept me company both then and for more than a decade before that.

Contents

Introduction

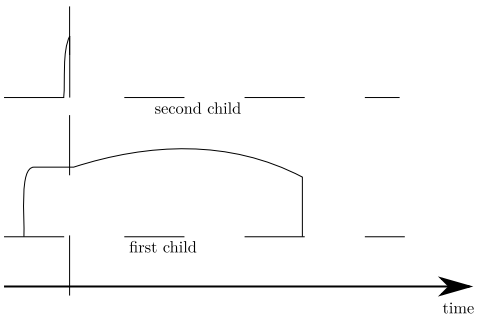
This thesis consists of six independent papers. Each of them is self-standing and can be read independently of the others.

Instead of duplicating the expository material contained in each of the six papers, this integrative introduction will provide a thematic overview of the thesis as a whole.

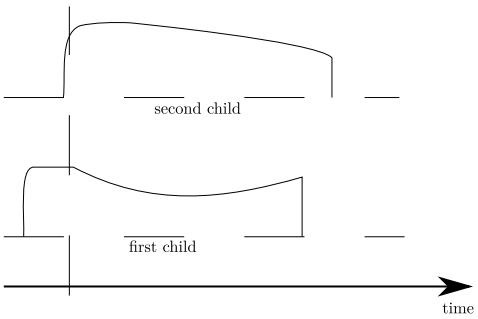
The central theme of this thesis is the unimportance of personal identity. Does it matter, for example, whether it is you or someone else who benefits from your current sacrifice? I take it that most people want to answer "Yes". I will argue that the right answer is often "No". I focus on examples from population ethics and deontological morality.

The issue of personal identity in population ethics can be illustrated nicely by an example of Broome's. He asks: "[w]hat resources should be used for saving young, perhaps premature, babies?" and imagines a situation where "a particular baby can be saved at the cost of reducing her sibling's standard of life" (2004: 9).

Following Broome, we can use the following diagram to illustrate this choice, where each possible person is allocated one horizontal axis and the height of the graph represents how well their life is going at any given time.

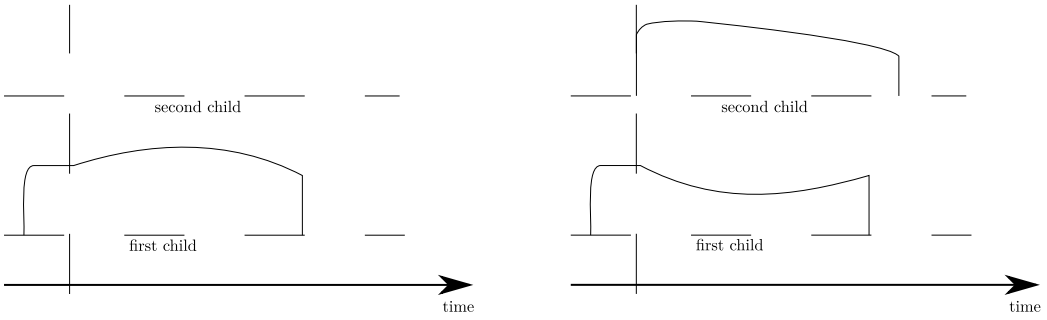


Help the first child



Save the second child

Now compare this case with one where the second child will only come to exist a little later.



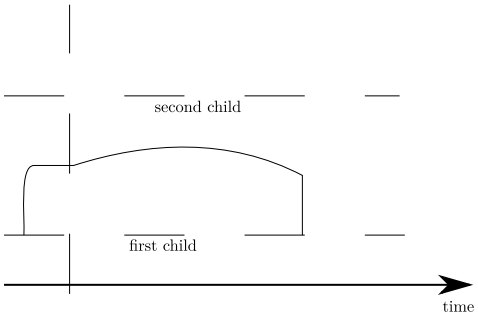
Help the first child Create a second child

As Broome says, the two cases differ

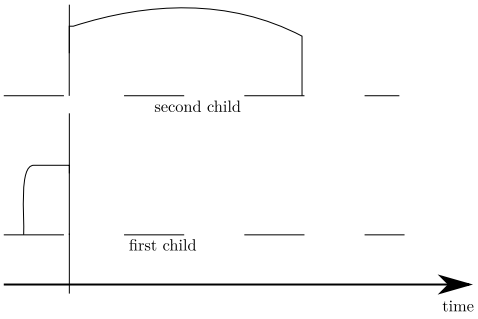
> "only in the second child's very short existence before the present. Many people would be inclined to think this small blip in the diagram makes all the difference to the problem. The value of prolonging the life of an existing person seems quite different from the value of creating a new person" (2004: 9-10).

If we share this intuition, we think that personal identity matters. We think it matters whether the second child has just come into existence or is about to.

For a perhaps starker example, imagine that, for the second child to exist, the first child's life not only has to be worse but has to end prematurely, as in the diagram below.



Save the first child



Create a second child

Here the only difference between the two courses of action is whether a given pattern of wellbeing is spread across two lives or just one. Many people think that this difference matters. This is another way in which personal identity might be important.

Many theories of population ethics agree that personal identity matters. For example, average utilitarianism, according to which the value of an outcome is the average wellbeing of people alive in it, implies that it is better to prolong lives rather than create new ones, at least if all lives at issue would be good.[1] This is also an implication of critical-level utilitarianism, according to which the value of an outcome is the total of everyone's welfare minus the positive critical-level parameter α.[2] Similarly, even when the number of people affected stays constant, the core egalitarian principle of Pigou-Dalton implies that it is better to spread out good years so as to equalize the quality of people's lives on the whole.[3] So,

[1] Compare Blackorby et al. (2005: 151-2).

[2] Critical-level utilitarianism is defended by Broome (2004) and Blackorby et al. (2005).

[3] A recent extensive discussion of the Pigou-Dalton principle is in Adler (2012: 339-356).

egalitarians also care about personal identity. There are many other examples, too.[4]

This sensitivity to personal identity has often been criticized. For example, Hudson's (1987) objection to average utilitarianism is that

> "there seems to be a graduated progression from clear-cut persons to clear-cut non-persons, rather than any sharp distinction. (…) It will make a moral difference whether a given history is considered to be the history of a single person or of two different but closely related persons. But distinctions of this sort are also matters of degree, so any such theory is unsatisfactory" (127).[5]

What are we to make of suggestions like these? Can we turn them into explicit arguments against theories of population ethics which attach importance to personal identity?

I take up this issue in the first two chapters of this thesis.

In Chapter 1: **"Intrinsic Concerns without Extended Selves"** I examine Derek Parfit's pioneering arguments for the moral unimportance of personal identity.[6] I try to find their best and most defensible versions.

First, I argue that almost any account of personal identity makes it implausibly extrinsic, both in cases of fission and superlongevity. This includes David Lewis's and Barry Dainton's accounts which are typically advertised as avoiding any

---

[4] They are detailed in the first chapter of my BPhil thesis "Metaphysics of Persons and Population Ethics".

[5] Similar comments can be found in Chapter 15 of Parfit's *Reasons and Persons*.

[6] See Parfit (1971, 1987, 2007).

extrinsicness in personal identity.[7] We no longer have to find out what happens in these sci-fi cases: all accounts of what happens lead to implausible extrinsicness. Since nothing of moral importance can be extrinsic in the way I argue personal identity is, it follows that personal identity doesn't matter. I reach a similar conclusion about cases of superlongevity which involve extremely long-lived people who are otherwise psychologically like us.

Second, I argue that cases of fission show that personal identity is subject to a special sort of indeterminacy, distinct from mere vagueness, but related to the sort of indeterminacy Hartry Field claimed to find in Newtonian physics.[8] I argue that personal identity is therefore not substantive and cannot be morally relevant.

While I think these Parfitian arguments likely work, it is important to take note of their limitations. They show, at best, that factors of moral importance, the presence of extra goodness or badness, say, cannot depend on how person-stages are packaged into people's lives. But this doesn't mean they cannot depend on other relations among people's person-stages. For example, we might still think that average welfare or equality of person-stages matters, in addition to the total of person-stage welfare in the world.

Chapter 2: **"Johnston *versus* Johnston"** discusses a different type of argument against the importance of personal identity, due to Mark Johnston. He argues that the intrinsicness of moral status combined with a four-dimensionalist ontology of David Lewis undermines any recognizably commonsense, person-based moral outlook.[9] Johnston suggests that the problem generalizes

---

[7] See Lewis (1983), Dainton (1992).

[8] See Field (1973).

[9] See Johnston (2016, 2017).

beyond four-dimensionalism to any broadly naturalistic theory of our place in the world.

While I am sympathetic to Johnston's challenge, I think it runs into some problems. For one thing, there are independent reasons to reject or revise the account of intrinsicness Johnston relies on. For another, his discussion assumes an overly simplistic connection between value and its metaphysical basis. To make a successful argument from metaphysics to ethics, we must not import unwarranted metaphysics into the ethics. I think my arguments in Chapter 1 have the virtue of being immune to that kind of criticism.

The next two chapters move on to discuss specific examples of sensitivity to personal identity in moral philosophy.

I begin with egalitarianism in Chapter 3: **"Egalitarianism and Population Size"**. Egalitarianism is well understood when it comes to situations where the number of people is fixed, but not when it comes to situations where creating new people is a possibility.

I argue that egalitarians face some uncomfortable choices in the latter type of situation. I pinpoint a form of egalitarianism that they have to accept if they want to stay true to the idea that relations between people matter but also to avoid some familiar objections, including the levelling-down objection, which alleges that egalitarianism absurdly implies there would be something good about blinding the sighted as a means of reducing inequality.[10]

I then go on to argue that this form of egalitarianism delivers implausible and unmotivated verdicts in cases where creating new people is a possibility. Some of

[10] This form of egalitarianism has recently been defended by Asheim and Zuber (2014). On the levelling-down objection, see Parfit (1991).

my examples are choices between prolonging a life and creating a new one. I conclude that egalitarians remain vulnerable to familiar objections to their view, chiefly the levelling-down objection.

Even if we give up on attaching importance to personal identity when it comes to determining good and bad, we might still think it is important when it comes to determining right and wrong. To see this, consider an example from Parfit:

> "We must decide whether to impose on some child some hardship. If we do, this will either (i) be for this child's own greater benefit in adult life, or (ii) be for the similar benefit of someone else – such as this child's younger brother. Does it matter morally whether (i) or (ii) is true?" (1987: 333).

Even if we think that the difference doesn't matter, in the sense that imposing a hardship is just as good in case (i) as in case (ii), we might think that it is nonetheless forbidden in the latter but permitted in the former. Personal identity makes a difference in permissibility, even if it doesn't make a difference in value. This is another way in which personal identity might be important.

This is the topic of Chapter 4: **"People in Suitcases"**. I argue that deontic constraints cannot be combined with the attractive idea of acting in everyone's interest. I show this by means of cases where agents need to make multiple decisions across time. I then argue that these problems force us to choose between orthodox deontology and broadly consequentialist views, and that we should opt for the latter. I also sketch a vindication of Harsanyi's use of the veil of ignorance. The chapter ends by suggesting that it is utilitarian consequentialism rather than standard deontology that can truly give importance to flesh-and-blood people rather than the abstract idea of personhood.

In the final two chapters I go back to some of the issues posed by Broome's initial example of the value of prolonging a life.

Recall that his example was meant to illustrate why we might care about whether a new life has just begun or is instead just about to begin. As Broome puts it in the passage cited above, "[m]any people would be inclined to think this small blip in the diagram makes all the difference to the problem" (2004: 10).

Note that this is true regardless of the size of the blip. If the blip is there at all, the example is about prolonging a life; if it isn't it, it is about creating a new life. We might think this is problematic as it appears to be a case of implausibly extreme sensitivity to arbitrarily small nonevaluative differences.

We might try to rule it out by appealing to the principle of hypersensitivity avoidance, according to which no two things can be arbitrarily close in nonevaluative terms but arbitrarily far apart in terms of value.[11]

Hypersensitivity can arise in multiple dimensions and follows from a wide range of views. The first dimension is time. This is essentially what is going on in Broome's example. Another dimension is personhood itself, as it is plausible that whether something is a person at a given time depends on mental and physical features which come in degrees.[12]

But it is not clear why we should accept hypersensitivity avoidance and what its implications exactly are. The last two chapters try to capture some of the intuitions behind hypersensitivity avoidance in a more principled and more familiar framework.

---

[11] I take this principle, with small changes, from Pummer (2019).

[12] This is the dimension that Pummer (2019) focuses on.

Chapter 5: **"Transfinitely Transitive Value"** introduces transfinite transitivity and transfinite acyclicity which generalize familiar value-theoretic principles of transitivity and acyclicity and applies these transfinite principles to the problem of evaluating outcomes where different numbers of people exist. I use these principles to argue that it is always good to add good lives, bad to add bad lives, and neutral to add neutral lives, where the value of a life is understood as the value for its subject. This conclusion rules out many prominent theories of how to compare outcomes where different numbers of people exist, such as average and critical level utilitarianism, and it allows us to reduce variable population axiology to fixed population axiology. Transfinite transitivity has a clearer motivation than hypersensitivity avoidance, it can play much of the same role in value theory, and it is also, in some respects, more liberal. The chapter also shows how these axiological conclusions can be used to support a limited version of the so-called procreative asymmetry.

Chapter 6: **"Transfinite Transitivity and Rational Choice"** aims to support transfinite transitivity by showcasing its role in the theory of rational choice. It argues that its role in compact option sets is broadly like that of transitivity in finite option sets. That role is to secure the existence of choiceworthy options, secure consistency of choice across different situations, and to protect agents from money pumps. The distinctions developed in this chapter are then applied to a classic infinite decision puzzle, Satan's Apple.

I hope that this thesis provides multiple viable routes, not necessarily driven by metaphysics, to a picture of morality where personal identity plays a lesser role than it does in much of our current moral thinking.

References

Adler, M. (2012). *Well-being and fair distribution*. Oxford: Oxford University Press.

Asheim, G. B., & Zuber, S. (2014). Escaping the repugnant conclusion: Rank-discounted utilitarianism with variable population. *Theoretical Economics, 9*(3), 629-650. doi:10.3982/TE1338

Bartha, P., Barker, J., & Hájek, A. (2014). Satan, Saint Peter and Saint Petersburg. *Synthese, 191*(4), 629-660. doi:10.1007/s11229-013-0379-9

Blackorby, C., Bossert, W., & Donaldson, D. J. (2005). *Population issues in social choice theory, welfare economics, and ethics*. Cambridge: Cambridge University Press.

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

Dainton, B. (1992). Time and division. *Ratio, 5*(2), 102-128.

Field, H. (1973). Theory change and the indeterminacy of reference. *The Journal of Philosophy, 70*(14), 462-481. doi:10.2307/2025110

Hudson, J. (1987). The diminishing marginal value of happy people. *Philosophical Studies, 51*(1), 123-137. doi:10.1007/BF00353967

Johnston, M. (2016). Personites, maximality and ontological trash. *Philosophical Perspectives, 30*(1), 198-228. doi:10.1111/phpe.12085

Johnston, M. (2017). The personite problem: Should practical reason be tabled? *Noûs, 51*(3), 617-644. doi:10.1111/nous.12159

Lewis, D. (1983). Survival and identity, with postscripts. In D. Lewis, *Philosophical papers, volume I* (pp. 55-77) Oxford University Press.

Parfit, D. (1971). Personal identity. *The Philosophical Review, 80*(1), 3. doi:10.2307/2184309

Parfit, D. (1987). *Reasons and persons* (2nd repr. with corrections of 1984 ed.). Oxford: Clarendon Press.

Parfit, D. (1991). *Equality or priority* University of Kansas, Department of Philosophy. Retrieved from http://hdl.handle.net/1808/12405

Parfit, D. (2007). Is personal identity what matters. *Ammonius Foundation,* Retrieved from http://www.stafforini.com/docs/parfit_-_is_personal_identity_what_matters.pdf

Pummer, T. (2018). Spectrum arguments and hypersensitivity. *Philosophical Studies, 175*(7), 1729-1744. doi:10.1007/s11098-017-0932-3

Chapter 1

Intrinsic Concerns without Extended Selves

Abstract: I defend two Parfitian arguments for the unimportance of personal identity in morality: the argument from extrinsicness and the argument from reductionism. The first starts by showing that personal identity is wildly extrinsic in cases where people divide and merge like amoebas, given any view of what happens in these cases. Given an additional plausible assumption about change over time, personal identity is wildly extrinsic in cases of extremely long-lived people whose bodies and minds gradually change like ours do. I conclude that nothing of moral importance can depend on personal identity, since morally important factors cannot be wildly extrinsic like that. Unlike Parfit's own arguments, mine assume very little by way of metaphysics of personal identity. Drawing on Field's and Lewis's discussion of indeterminacy in theoretical identifications, my second argument identifies a coherent and plausible sense in which personal identity is not substantive and, hence, unimportant. My two arguments are meant to avoid multiple objections to Parfit, due to Ernest Sosa, Mark Johnston, David Lewis, Barry Dainton, and Anthony Brueckner, among others. I end by suggesting what my arguments mean for value theory.[1]

Word count: 11956

---

[1] I would like to thank Ralf Bader, Teru Thomas, Michal Masny, Tomi Francis, and Jacob Trefethen. Thanks also to the audiences at the 2018 ISUS conference in Karlsruhe and at the 2018 CEPPA conference in St Andrews where my commentator was Quan Nguyen.

1 Introduction

Parfit gave two kinds of arguments for his notorious claim that personal identity is not what matters.

> **Argument from Extrinsicness**. Personal identity is an extrinsic matter. What matters is never like that. So, personal identity is not what matters.[1]
>
> **Argument from Reductionism**. There is, at least sometimes, no deep fact of the matter about personal identity. What matters is never like that. So, personal identity is not what matters.[2]

Both arguments face serious objections, however.

One objection to the first argument is that personal identity is, in fact, not an extrinsic matter. Parfit got his metaphysics wrong.[3] Another objection is that many important things in life depend on extrinsic factors. So, it seems, Parfit got his ethics wrong, too.[4] And a key objection to the second argument is that there is no sense to be made of "deep fact of the matter" which puts personal identity on the shallow, nonfactual side, but keeps much else on the other side.[5] It also looks like we cannot hold both arguments at the same time, as Parfit seems to want to do. After all, how can we conclude that there is no fact of the matter about personal identity but also that personal identity is definitely extrinsic?[6]

---

[1] See Parfit (1971, 1987: 253-273, 1993). My formulation is close to Parfit's own in his (2007).

[2] See Parfit (1971, 1973, 1982, 1987: 219-243, 307-347, 1995). The 1995 paper is reprinted virtually unchanged as Parfit (2011) and appears as the initial sections of Parfit (2007).

[3] See Perry (1972), Lewis (1976), Noonan (1985), Dainton (1992, 2008: 364-408).

[4] See Sosa (1990).

[5] See Shoemaker (1985), Sosa (1990), Johnston (1992, 1997), Garrett (1992, 1998: 83-94).

[6] See Brueckner (1993).

In this paper I will try to make Parfit's arguments work. The arguments I will end up defending aren't Parfit's own. But, I hope, he would have found them congenial.

In section 2 I show that, in cases where a person divides like an amoeba, personal identity is an extrinsic matter on any view what happens in these cases and that what matters cannot be extrinsic in the same way. In section 3 I argue that a similar argument goes through in cases of people who are extremely long-lived but otherwise like us. In section 4 I show how to make sense of Parfit's argument from reductionism. Unlike Parfit's arguments, my two arguments are compatible. If they are sound, personal identity is not what matters. In section 5 I explain what this means for value theory.

2 Extrinsicness and fission

Nuclear fission is the splitting of one atom into two. In the case of persons, fission is the splitting of a person's body and mind into two.

As an example consider

> **Double Transplant**. Derek's body is fatally injured. His brain is divided in half, and each half is successfully transplanted into two healthy but brainless bodies of his twin brothers. Each of the resulting people believes that he is Derek, remembers living Derek's life, and so on. And he has a body that is very like Derek's.[7]

What happens in this case? Does Derek survive as the left-brained person, Lefty? Or as the right-brained person, Righty? Or maybe as both? Or neither? Or something else entirely?

In metaphysical terms, the basic problem of fission is the conflict between our usual ways of identifying people at a time and across time.[8] Typically, we think that, at any given time, there is one person per mind and body, and one mind and body per person. In Double Transplant that would mean that there is one person before the operation and two people afterwards. We also typically think that having enough of someone's brain and body is sufficient for being that person. But in Double Transplant that would mean that the single pre-op person is both post-op people, contradicting basic logic of identity.

---

[7] This is a version Parfit's (1987: 254) case of My Division. Fission cases have been discussed before by Williams (1956), Prior (1957), Wiggins (1967: 43-58), and even earlier by Locke, Clarke, Priestley, and Hazlitt – see Martin, Barresi and Giovannelli (1998).

[8] This point is clearly articulated by Dainton (1992: 103). See also Johnston's (1989a) and Sattig's (2015: 104-133) lists of principles at stake in fission.

There are four main ways to resolve that tension.

Termination View. Derek survives as neither Lefty nor Righty.[9]

Asymmetric View. Derek survives as Lefty or as Righty but not both.[10]

Multilocation View. Derek survives as both Lefty and Righty.[11]

Cohabitation View. There are two people all along, Lefty and Righty.[12]

What does fission show us about what matters?

2.1 Parfit's Argument from Fission

Parfit took fission to show that personal identity is not what matters in the special sense that it isn't what justifies self-interested concern that we typically have towards ourselves in the future.[13]

It is natural to read Parfit's argument here as follows.[14] First, the true view about fission is either the termination view or the asymmetric view. So, in Double Transplant, Derek is either not Lefty or not Righty, or neither Lefty nor Righty.

---

[9] See Nozick (1981: 29-70), Garrett (1990, 1998: 58-70), Unger (1990: 255-294).

[10] See Swinburne (1974), but also Bader (forthcoming) who thinks that, determinately, Derek is either Lefty or else Righty, but it is indeterminate which one.

[11] Dainton (1992, 2008: 364-408). A recent convert is Johnston (2010: 308), although he defended the bare coherence of multilocation already in his (1989a).

[12] Lewis (1976), Robinson (1985), Heller (1987), Langford (2007). Noonan (1985, 2019: 233-250) and Perry (1972) defend similar views.

[13] Parfit (2007) makes it clear that this concern is to be understood in terms of anticipation.

[14] That is how Parfit's argument is presented in Noonan's (2019) and Garrett's (1998) textbooks. But Parfit doesn't actually accept it, as he makes clear in his (1993) reply to Brueckner (1993). Instead, he thinks there is no fact of the matter about personal identity in fission, and the termination view is just the best proposal about how that case should be described.

Still, the relationship between Derek and Lefty in Double Transplant is intrinsically the same as the relationship between Derek and Lefty in the case of

> **Single Transplant**. Same as Double Transplant except that the right half of Derek's brain is destroyed on the way to the operating room, and only the left half is successfully transplanted.

But whether one is justified in having self-interested concern towards some person in the future is, plausibly, an intrinsic matter. So, since Derek is justified in having self-interested concern towards Lefty in Single Transplant, he is also justified in having it towards Lefty in Double Transplant. And, by an analogous argument, he is also justified in having it towards Righty in Double Transplant.

But, we are assuming, Derek cannot be both Lefty and Righty. So, it must be that he has justified self-interested concern towards someone who is not him. So, it isn't personal identity that justifies that sort of concern. Parfit thinks this claim has revisionary implications for rationality and morality.[15]

But to make his argument, we would first have to rule out multilocation and cohabitation views of fission. But why? Isn't it better to tweak the metaphysics rather than the ethics?

2.2 My Argument from Fission

Since I sympathize with that objection to Parfit's argument, I want to argue that personal identity is an extrinsic matter, in a particularly implausible way, whatever view of fission we take, provided we agree with common sense in

[15] See Parfit (1987: 307-347). I take it that it is this last step that writers like Jeske (1993) and Shoemaker (2002) want to resist.

ordinary, non-fission cases.[16] And instead of starting with the narrow question of what justifies self-interested concern, I will argue, more directly, that morally relevant factors of all sorts cannot depend on personal identity. As an example, I will focus on the principle that

> (†) it is in some way bad when, through no fault of their own, people suffer burdens without enjoying counterbalancing benefits.

I will often shorten this to "it is bad when people suffer uncompensated burdens". I hope no confusion will result from this abbreviation.

I take it that this is one plausible way to cash out the claim that personal identity is what matters. It can help explain, for example, the difference many people see between the following two cases.[17]

> **Bone Autograft**. Derek and his twin brother are unconscious after an accident. Derek's bones are badly broken while his twin brother has no serious injuries. The doctors help Derek by grafting some of his remaining healthy bone tissue.
>
> **Bone Allograft**. Same as Bone Autograft except that the doctors help Derek by grafting some of his brother's healthy bone tissue.

In Bone Allograft Derek's brother undergoes a sacrifice for Derek's benefit. In Bone Autograft it is Derek himself who undergoes that sacrifice. If it is bad when people suffer uncompensated burdens, there is extra badness in Bone Allograft that is absent in Bone Autograft. I will argue that the principle that it is bad

---

[16] So, I agree with Ross (2014) that revising our metaphysics won't solve Parfit's problem. But, insofar I understand Ross's paper, there is no overlap in our arguments for that claim.

[17] Compare, for example, Nozick (1974: 32-33).

when people suffer uncompensated burdens leads to problems on any view of what happens in fission.

But even though I focus on that principle, I think my arguments work in the case of other morally relevant factors.[18] To that extent they support the general claim that personal identity is not what matters.

Let's start with asymmetric and termination views. They both imply that Derek is either not Lefty or not Righty. Let's assume, harmlessly, that he isn't Lefty.

Now consider the following two cases.

> **Case 1**. Same as Double Transplant except that, before the operation, Derek undergoes a sacrifice for the sake of Lefty's later benefit.
>
> **Case 2**. Same as Single Transplant except, before the operation, Derek undergoes a sacrifice for the sake of Lefty's later benefit.

We can represent them as follows.

[18] For example, I think they work against principles like "it is wrong to kill people", "it is wrong to break promises to people", "it is bad if some people are worse-off than others through no fault of their own", "it is bad if people are punished for crimes they didn't commit".

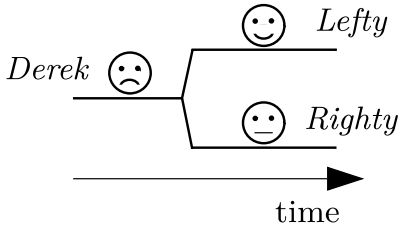


Figure 1-1. Case 1

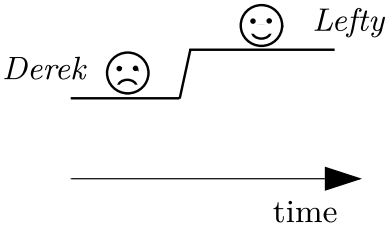


Figure 1-2. Case 2

We are assuming that Derek isn't Lefty in Case 1. But Derek is obviously Lefty in Case 2.

So, according to asymmetric and termination views, whether Derek is around to reap the fruits of his earlier sacrifice depends on whether Derek's other hemisphere is successfully transplanted.

But that might depend on what is happening very far away from the left hemisphere. For example, imagine that the right hemisphere is shipped to a treatment facility on Mars while the left hemisphere stays here on Earth. Immediately afterwards all contact with Mars is lost. To find out if Derek really survived the transplant operation on Earth, we might have to fly to Mars.

And, so, if it is bad when people suffer uncompensated burdens, we might also have to fly to Mars to find out whether there is something bad about Derek's sacrifice for Lefty's sake here on Earth. That is implausible.

We can put the point by saying that badness cannot be *wildly extrinsic*: it cannot non-causally depend on far-away goings-on in the way imagined.[19]

---

[19] I will say more about wild extrinsicness below. Wasserman (2005) uses "wildly extrinsic" in a similar sense. I don't think I need to presuppose any particular account of the intrinsic/extrinsic distinction to run my argument. For some options, see Lewis and Langton (1998) and Bader

Wildly extrinsic badness threatens partly because asymmetric and termination views of fission make personal identity over time into something wildly extrinsic. They both imply that whether a successful hemisphere transplant guarantees one's own survival depends on what is going on with one's other hemisphere, possibly far away.[20]

We might doubt my claim that badness cannot be wildly extrinsic. After all, aren't there many examples of extrinsic value?

Firstly, and most obviously, there are cases of *instrumental value*. For example, a visit to a dentist might be painful and, so, bad when considered in itself, but extremely good, depending on what does or could happen many years in the future.

Next are examples of *constitutive value*. An event can be good or bad depending on how it contributes to a broader pattern of value to which it belongs. For example, giving an extra benefit to someone well-off in Europe might exacerbate inequality between people of the world as a whole and, hence, it might be not very good or even bad.

Lastly, we have putative examples of *final extrinsic value*, things that are good or bad in themselves but because of how they relate to something external. Kagan's (1998) example is "the pen used by Abraham Lincoln to sign the Emancipation Proclamation, freeing the slaves":

(2013). In the context of personal identity, intrinsicness is typically discussed under the heading of the so-called only *a* and *b* principle, due to Wiggins (1980: 96). See Johnston (1989a) and Hawley (2005).

[20] This is well-known, see, for example, Heller (1987), Noonan (1985), Garrett (1990), Hawley (2005).

> "the pen's defining instrumental moment is now long since over. But by virtue of that history, we might say, it now possesses intrinsic value: it is something we could reasonably value for its own sake. The world is the richer for the existence of the pen" (285).[21]

But the sort of extrinsicness we see in instrumental, constitutive and final extrinsic value is quite unlike the extrinsicness I criticized in cases of fission and fusion, where wild extrinsicness in personal identity is not necessarily accompanied by any of the backdrop needed for plausible instances of extrinsic value.

To see this, assume – for the sake of illustration – that Derek is neither Lefty nor Righty in Double Transplant. If people's uncompensated suffering is bad, then there is extra badness on Earth if Derek's other hemisphere is successfully transplanted somewhere else, maybe on Mars. But Derek's sacrifice before the operation doesn't necessarily have any effects, good or bad, on what's happening on Mars. We are also free to assume that if and when Righty wakes up, he won't have any painful memories of Derek's sacrifice on Earth. And Derek's sacrifice here on Earth doesn't necessarily spoil any broader evaluative pattern either. Lastly, the extra badness of Derek's sacrifice cannot be explained in terms of its past or potential effects, like the extra value of Lincoln's pen which is, arguably, due to its beneficial past effects.[22]

---

[21] Other examples can be found in Korsgaard (1983) and Rabinowicz and Rønnow-Rasmussen (2000).

[22] Value of rarity provides a more problematic example, as it is, on the face of it, wildly extrinsic. But, for that reason, it is implausible. As Beardsley (1965) points out, if rarity was valuable, a thing here might become more valuable just because other things of its kind are destroyed in some

So, I don't think that appealing to more familiar forms of extrinsic value will help those who think people's uncompensated suffering is, in some way, bad.

Still, we might think that the problem goes away once we move to cohabitation and multilocation views of fission, often advertised as avoiding any extrinsicness in personal identity. But that isn't so.

To see how the problem arises for the cohabitation view, consider the following two cases.

> **Case 3**. Same as Double Transplant except that Derek undergoes a pointless sacrifice before the operation.
>
> **Case 4**. Same as Single Transplant except that Derek undergoes a pointless sacrifice before the operation.

We can represent them as follows.

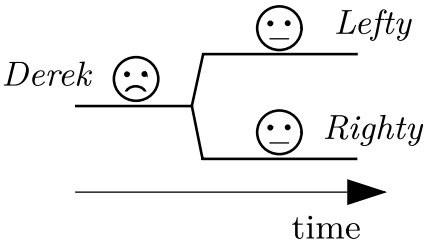


Figure 1-3. Case 3

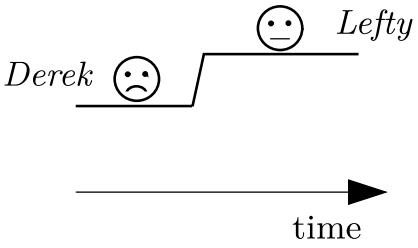


Figure 1-4. Case 4

According to the cohabitation view, Derek isn't really a single person before the operation in Case 3. There are two people, Lefty and Righty, sharing Derek's body and mind all along. "Derek" is really an ambiguous name. But, in Case 4,

---

remote corner of the universe. There are also ways to account for intuitions about cases like rarity without making value wildly extrinsic. See Hurka (1998) and Regan (2003).

there is just one person throughout, Lefty, even though, on the day before the operation, the world looks just the same in both cases.

So, whether there are two miserable people before the operation or just one depends on whether that future operation succeeds.

But that might not be known for many years to come. Perhaps no one can even imagine that such an operation will ever take place. And, obviously, the success of the operation has no causal influence on what is going on with Derek's body many years before.

So, if it is bad when people suffer uncompensated burdens, there are two instances of that badness if the future operation succeeds but just one if it does not. Badness is again wildly extrinsic. That is implausible.

To make this vivid, imagine that you are standing next to someone you have known as Derek his whole life. He is miserable. But many years from now his hemispheres will be separated and transplanted into two other bodies. It is hard to believe that the badness of what is right in front of us depends on the future like that.[23]

We might think we could solve this problem for the cohabitation view by joining Lewis (1976) in counting by identity-at-a-time rather than identity.

That is, we might say that two people are identical-at-a-time if they share the same body and mind at that time. We can still say they aren't identical full stop if they don't *always* share the same body and mind. So, while there are two

[23] Nolan and Briggs (2015) and Campbell (ms) make a related point that the cohabitation view leads to double-counting temporal wellbeing. They don't realize that this is an example of wild extrinsicness to which other views of fission also lead.

miserable people before the operation, when counting by identity, there is just one miserable person, when counting by identity-at-that-time.

Counting by identity-at-a-time does help, for example, when we are asking, on the day of the operation, how many people came into the operating room. It delivers the natural answer "one". But it doesn't help when we take a more timeless perspective, as we naturally do when thinking about what's good or bad about a situation as a whole. In that context even Lewis tells us to "count by identity, if we count from the standpoint of no definite time".[24]

What about the multilocation view? According to that view, Lefty and Righty are really the same person after the operation. But how?

There are a few different ways to make sense of that basic idea. One is to say, following Johnston (1989a), that Derek survives spatially separated from himself in the fashion of a universal. Another one, following Dainton (1992), is to say that Derek will be in two places at the same time much like a time-traveller meeting his older self. Johnston (2010: 308) now subscribes to a different model, according to which Derek is really a higher-order individual, much like the species tiger, and, so, can survive in two places at the same time, much like the tiger itself survives both in Bengal and Sumatra.

To see how the problem arises for all these versions, consider the

> **Short-Line Case**. Same as Single Transplant except that, immediately after the right hemisphere is destroyed, its perfect replica appears out of thin air and is then successfully transplanted. Each of the resulting people

[24] Nolan and Briggs (2015: 403) make a similar point. I think this also shows that stage theory doesn't help either. This is because it arguably has to revert to something like Lewis's timeless perspective in the context at hand. See Sider (1996: 448).

believes that he is Derek, seems to remember living Derek's life, and so on.[25]

In this case, Righty's post-op existence is like a shortline railroad: intrinsically, it could belong to the bigger system with Derek and Lefty, but, as it happens, it doesn't.

And now consider the following two cases, one based on Double Transplant, the other on the Short-Line Case.

Case 5. Same as in Double Transplant except that, after the operation, Righty undergoes a sacrifice for the sake of Lefty's simultaneous benefit.

Case 6. Same as in Short-Line Case except that, after the operation, Righty undergoes a sacrifice for the sake of Lefty's simultaneous benefit.

We can represent them as follows.

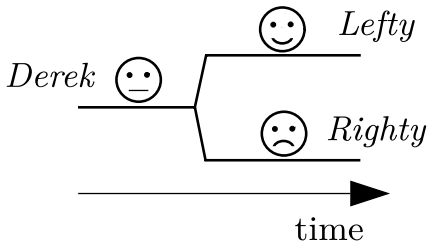


Figure 1-5. Case 5

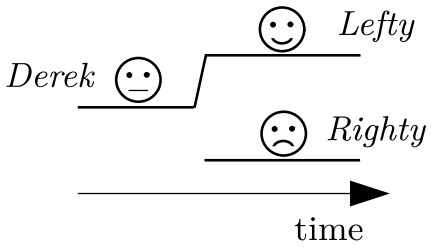


Figure 1-6. Case 6

According to the multilocation view, Derek is both Lefty and Righty in Case 5. But, plausibly, he is just Lefty in Case 6. Righty is just an accidental replica of pre-op Derek.

[25] Compare Parfit's (1987: 201) Branch-Line Case.

So, whether Righty undergoes a sacrifice for his own sake or for someone else's sake depends on the fine details of the operation, possibly long time ago. Lefty and Righty might now live far away from each other and as different from each other as any two people: you and me, for example.

But, if it is bad when people suffer uncompensated burdens, we would need to find out what happened long time ago to establish whether it is bad if now Righty takes a hit for Lefty. That is implausible.

To make this vivid, imagine, for example, that Derek's operation took place years ago, but we aren't sure whether it was Derek's right hemisphere that ended up transplanted or a mere replica. Yesterday Lefty and Righty, who are now perfect strangers to each other, were involved in a serious accident. To fix Lefty's many broken bones, we need to painfully extract some bone tissue from Righty. Would it be good or bad if we did that? It is hard to believe the answer could depend on the far past in the way imagined.

The problem arises because the multilocation view makes personal identity at a time implausibly extrinsic. In that respect, multilocation is really not that different from cohabitation. For example, Dainton's (2008) telling objection against Lewis's cohabitation proposal, applies, with little change, to his own multilocation proposal:

> "Although Lewis allows us to retain the doctrine that streams of consciousness are reliable guides to the persistence of subjects, this comes at a cost: consciousness is no longer a reliable guide to the *existence* of subjects. It is natural to think that a single stream of consciousness necessarily belongs to a single persisting subject, but if Lewis is right, this isn't the case: there is no limit on the number of

subjects to whom your current state of consciousness might belong…" (375).

We now see that the multilocation view also implies that consciousness isn't a reliable guide to the existence of subjects, and that there is no limit on how many states of consciousness can belong to a single subject.[26]

Still, friends of multilocation might object that although fission shows personal identity at a time to be an extrinsic matter, it doesn't show it to be *wildly* extrinsic. After all, in Double Transplant but not in the Short-Line Case, Lefty and Righty share a common cause in the form of Derek. So, there is some causal explanation of extrinsic variation in personal identity.

Fair enough. We can find a truer case of wild extrinsicness in personal identity if we consider fusion: the time-reversed image of fission. Consider, for example,

> **Double Graft**. Two twin brothers are badly injured. They only have one healthy brain between the two of them. Their brains are successfully transplanted into a healthy but brainless body of their third brother. Because they spent a lot of time together, the two brothers are very psychologically similar before the operation. The resulting person

[26] Dainton might reply that, in cases of time travel, anyone has to deal with extrinsicness in personal identity at a given time. For example, imagine that, in 1984, Derek meets someone who looks like his older self travelling back in time from 2011. Whether Derek meets himself rather than an impostor depends on whether time travel machines are actually going to be invented in 2011. So, the number of people present in 1984 depends on what happens in 2011. But that isn't *wild* extrinsicness because whether Derek meets himself in 1984 depends on causal connections between Derek in 1984 and the purported time-traveller. Multilocation in fission cases wouldn't be like that.

remembers living their lives, shares their intentions, and so on. And he has a body that is very like theirs.

Contrast this case with

> **Single Graft**. Same as Double Graft except that the operation fails and one of the brothers dies on the operating table.

I will use "Lefty" and "Righty" as short for "the left-brained person" and "the right-brained person", as before. The resulting person I will call "Derek". Now consider the following two cases.

> **Case 7**. Same as Double Graft except that, before the operation, Righty undergoes a sacrifice for the sake of Lefty's simultaneous benefit.

> **Case 8**. Same as Single Graft except that, before the operation, Righty undergoes a sacrifice for the sake of Lefty's simultaneous benefit.

We can represent them as follows.

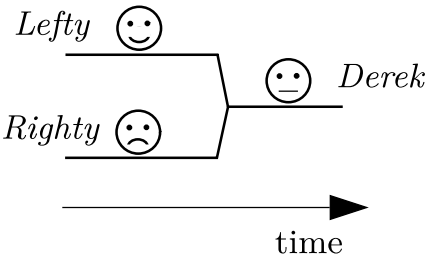


Figure 1-7. Case 7

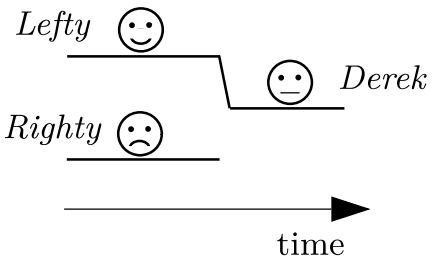


Figure 1-8. Case 8

Both fission and fusion show a conflict between our usual ways of identifying people at a time and across time. If we resolve that conflict one way when time flows forward, why shouldn't we resolve it the same way when time's arrow is

reversed? So, it is plausible to think that if fission involves multilocation, then so does fusion.[27]

But now I can run pretty much the same argument here as with Cases 5 and 6 before. This time whether Lefty is the same person as Righty before the operation depends on whether the future operation succeeds. And, obviously, the success of that operation has no causal influence on what is going on with Derek's body many years before.

So, whether Righty undergoes sacrifice for his own sake or for the sake of someone else depends on what happens in that future. And, so, it has no causal explanation. But if it is bad when people suffer uncompensated burdens, that means that badness is wildly extrinsic. That is implausible.

I conclude that no account of fission can escape wild extrinsicness in personal identity. If the presence or absence of a morally relevant factor depends on personal identity, that factor will also be wildly extrinsic. Whatever we think about this sort of extrinsicness in personal identity, it is definitely not plausible in what matters. So, we should reject principles that make what matters dependent on personal identity.

In particular, I suggest we should reject the principle that it is bad when people suffer uncompensated burdens. In general, we should think that no goodness or badness can depend on how person-stages are packaged into lives. However, it is important to note that this still allows goodness and badness to depend on other

[27] Dainton (2008: 400-405) agrees. Parfit (1971) and Lewis (1976) also treat fission and fusion symmetrically. The only exception I know of is Hawley (2005) who thinks fission cannot be fatal for the pre-op person, even if fusion is.

mental and physical relations among person-stages, as I will explain in more detail below.

2.3 Broome's Argument from Replacement

It is finally worth comparing my argument from fission with a related argument, proposed by Broome (1991a, 1991b: 230-237, 2004: 221-223).

Broome appeals not to fission and fusion but to cases like the following two.

> **Midlife Voyage**. You spend the first 50 years of your 100-year-long life in Europe and the rest in America, making a transatlantic trip on your 50th birthday. Your last years in Europe are full of hardship and suffering. Your move to America brings success and happiness.
>
> **Midlife Replacement**. Same as Midlife Replacement except for the following details. On your midlife transatlantic voyage your body is instantaneously annihilated. Then, completely randomly, a replica of you freakishly materializes out of thin air and takes your place. The replica lives out the remaining 50 years of your life just as you would.

We can represent these cases as follows.

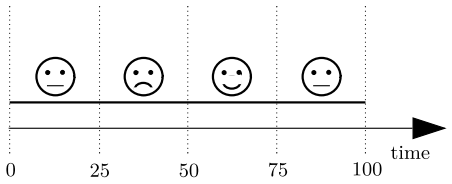


Figure 1-9. Midlife Voyage

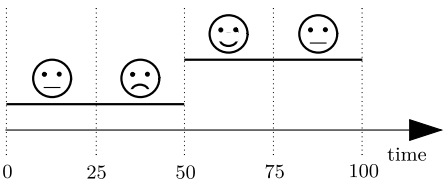


Figure 1-10. Midlife Replacement

Causal influence is, arguably, necessary for personal identity over time. So, in Broome's Midlife Replacement, your replica isn't you, even though your replica's American life is intrinsically like yours would be.

From cases like these Broome wants to conclude that there is no extra goodness or badness due to how person-stages are packaged into people's lives. His argument starts from the premise that

> "if the unifying relations between a person's stages did not hold, but the good and bad of each stage was just as it actually is, then the good and bad in the world would be just as it actually is" (1991b: 234).

By "unifying relations" Broome means relations of causally mediated mental and physical similarity that are typically taken to be the basis of personal identity over time.

Midlife Replacement only differs from Midlife Voyage in whether these relations hold across the Atlantic. So, Broome's principle seems to imply, for example, that there can be no extra goodness or badness due to the fact that you reap the fruits of your earlier sacrifices in Midlife Voyage but not in Midlife Replacement.

Broome's argument runs into a number of serious problems, however. The first is that, as Broome is well-aware, it might be impossible to imagine away the unifying relations while preserving the good and bad of each stage. So, we might doubt

whether your replica's life is Midlife Replacement is as good as yours would have been in Midlife Voyage.

But the second, more serious problem is that Broome's starting premise is implausible. It says not only that personal identity over time is irrelevant to goodness and badness, but also that the unifying relations (such as mental and physical continuity) are likewise irrelevant. That's too strong. At the corresponding place in my argument, I appeal to the idea that badness cannot be wildly extrinsic. I think that's a better place to start.

3 Extrinsicness and superlongevity

In addition to cases of fission and fusion, Parfit (1971, 1987: 303-5) also appeals to cases of superlongevity, involving people with "everlasting bodies, which gradually change in appearance", and whose psychologies gradually change, too. Indeed, in (1976: 89), he says that that sort of case is more important than cases of fission. But I don't think Parfit makes very good use of cases of superlongevity. In this section I will try to do better. I will show that they are arguably cases of wild extrinsicness in personal identity and, so, can be used to the same effect as cases of fission and fusion.[28]

As a concrete example, let's work with

> **Methuselah's Case**. Methuselah dies at age 969. At age 100 he still remembers his childhood. But at age 150 he has hardly any memories that go back beyond his 20th year. And at age 200 he has hardly any memories that go back beyond his 70th year. And so on. By his 200th year there is almost no trace of his opinions and character at age 70, but his opinions and character at age 200 also vanish almost without trace by his 330th year. And so on.[29]

We might doubt whether a single person can persist through Methuselah's entire career. As Lewis (1976: 17) puts it, for a person to persist "change should be gradual rather than sudden, and (at least in some respects) there should not be too much change overall". We can elevate Lewis's second clause into

---

[28] Neither Parfit (1971, 1987) nor Lewis (1976) see cases like Methuselah in terms of extrinsicness like I do. Rather they use them to make points about the graded character of self-interested concern.

[29] This is a version of Lewis's (1976) case.

> **The "Not Too Much Change" Proviso**. If a person at one time is sufficiently dissimilar in relevant respects from a person at a different time, they cannot be one and the same person.

I think the best way to see the problem posed by Methuselah's Case is as showing a conflict between Lewis's proviso and the desire to avoid wild extrinsicness in personal identity. I will give some reasons for the proviso later on, after I explain how the problem arises.

For simplicity, let's follow Lewis in assuming that there is too much change over 137-year-long stretches of Methuselah's career, but little enough change otherwise. So, Methuselah's career cannot belong to a single person. At most its 137-year-long segments can. But which ones?

We have two main options. The first one is the

> **Separation View**. Nonoverlapping 137-year-long segments of Methuselah's career belong to distinct persisting people.

It is natural to count them off from the beginning, so that the first 137-year-long segment belongs to a single person, born in year 1, dead in year 137. Then the next 137-year-long segment belongs to another person, born in year 138, dead in year 275. And so on.[30]

The second option is the

> **Overlap View**. Every 137-year-long segment of Methuselah's career, whether overlapping or not, belongs to some one persisting person.

[30] My arguments would work with small changes if we instead counted off from the end of Methuselah's career or from the middle.

So, for example, at his 100th birthday Methuselah's body is shared by at least 100 people: 99 people that came to be at Methuselah's previous birthdays and one person who comes to be at his 100th. If time is dense, so that there is a distinct time between any two times, there are infinitely many people at each of Methuselah's birthdays.[31]

Both options lead to wild extrinsicness in personal identity. And we can parlay that into problems for the claim that personal identity is what matters, focussing once again on the principle that it is bad when people suffer uncompensated burdens.

The following two cases illustrate what the problem is for the separation view.

> **Timely Methuselah**. Same as Methuselah's Case except that Methuselah's 137th year is full of suffering and anguish, while his 138th year is full of equal happiness and peace.
>
> **Delayed Methuselah**. Methuselah springs into existence fully formed as he was at age 1 in Timely Methuselah. His career is otherwise the same.[32]

We can represent them as follows.

---

[31] As far as I can tell, one's views about fission and superlongevity can be relatively independent. For example, there might be cohabitation in cases of fission but not in cases of superlongevity.

[32] Some details might have to be added here to get around the necessity of origin. For example, we might have to add that Methuselah comes from the same gametes in both cases.

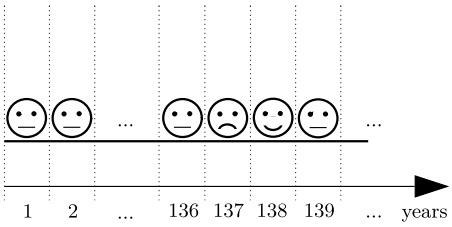


Figure 1-11. Timely Methuselah

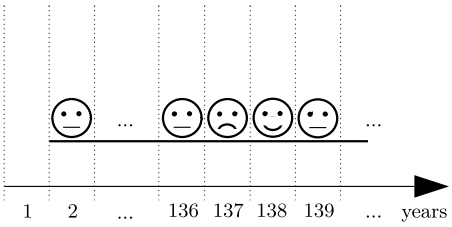


Figure 1-12. Delayed Methuselah

In Timely Methuselah, we divide Methuselah's career into eight 137-year-long lives and one 10-year-long one at the very end. In Delayed Methuselah we make similar divisions, except shifted by a year. So, in that case we count years 137 and 138 as belonging to some one persisting person, with the next person born in year 139.

So, whether there is a single person who goes through both the hardships of the 137th year and the rewards of the 138th depends on events that took place more than 100 years before. What's more we are assuming that these events had no effect on the later course of his career.

If it is bad when people suffer uncompensated burdens, then there is extra badness in Timely Methuselah that is missing in Delayed Methuselah. The presence of that badness is wildly extrinsic.

We can make the problem more pressing. First: we can focus instead on Methuselah's 959th and 960th years. Whether a single person lives through both of them depends on events that took place almost a millennium before. Second: we can instead look at Methuselah's career in terms of weeks, days, seconds, and so on, rather than years.

It should by now be clear that the overlap view doesn't succeed where the separation view fails. To see this, consider the following two cases.

Curtailed Methuselah. Same as Methuselah's Case except that Methuselah dies just before his 137th birthday and his 2nd year is full of suffering and anguish.

Expanded Methuselah. Same as Curtailed Methuselah except that Methuselah dies a year later, just before his 138th birthday.

We can represent them as follows.

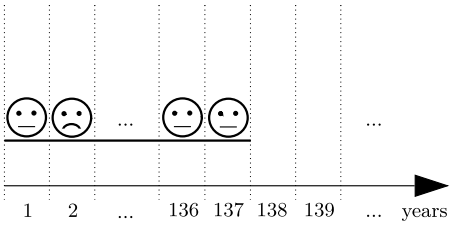


Figure 1-13. Curtailed Methuselah

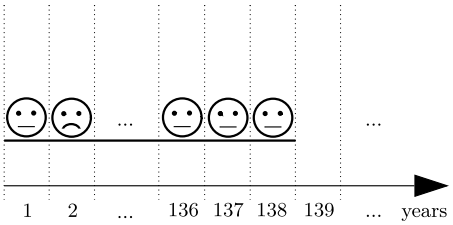


Figure 1-14. Expanded Methuselah

In Curtailed Methuselah, Methuselah's career is exactly 137 years. So, Lewis's "not too much change" proviso doesn't kick in: there is just one person throughout. But it does kick in in Expanded Methuselah. So, according to the overlap view, that case is an example of massive overpopulation.

So, whether Methuselah's body and mind are shared by one or many more people depends on whether his overall career is just shy of 137 years or a little longer. And whether there is one or many more miserable people living through Methuselah's 2nd year also depends on that.

So, if it is bad that people suffer uncompensated burdens, there are many more instances of that badness in Expanded Methuselah than in Curtailed Methuselah. Since there is no backwards causation going on, the presence of that badness is therefore wildly extrinsic. Again, we can make the problem more pressing by running the whole argument in terms of weeks, days, seconds, and so on.

So, much like cases of fission and fusion, cases of superlongevity show that we must allow for wild extrinsicness in personal identity. Since what matters cannot be like that, it follows that personal identity is not what matters.

But all this depends on granting Lewis's "not too much change" proviso. Why should we believe it? All that Lewis has to say in its defence is that, unlike for us, for Methuselah,

> "the fading-out of personal identity looms large as a fact of life. It is incumbent on us to make it literally true that he will be a different person after one and one-half centuries or so" (1976: 30).

So, it is no wonder that Johnston (1989a: 387, 2017: 642) dismisses Lewis's proviso as not well-motivated.[33] Unfairly, I think. We seem to accept a "not too much change" proviso in other cases besides superlongevity.

First, as Williamson (2013 [1990]) points out and as fans of Mr Squiggle have known long before, the creation of drawing can go through many gradual stages, often changing the initial sketch beyond recognition:

> "A panorama of the courtroom in which no witness appeared could eventually become a study of a witness's face. (…) The illustrator could carry out her editor's instruction to do a new drawing by making such a series of changes, if the result were far enough from the starting point" (136).

This is a "not too much change" proviso at work. It is natural to think languages work that way, too. To use another example of Williamson's, Latin and Italian

---

[33] Johnston (2017) thinks Lewis's discussion of Methuselah points to his own, more troubling problem involving *personites*, person-like but shorter-lived continuants. As I argue in chapter 2 the opposite is true.

are different languages, even though, we might assume, the change of one into the other was gradual enough. Biological species provide another relevant example.[34]

Lastly, we have modality: it is natural to think that an object would have existed even if its constitution was somewhat different but not if it was completely different. For example, a given wooden table could have existed if partly made of ice but not if wholly made of ice. This is a "not too much change" proviso but across the modal, rather than temporal, dimension.[35]

[34] Similar examples are engagingly discussed by Van Deemter (2010: 19-30, 54-69).

[35] This example is relevant to so-called Chisholm's Paradox. See Chisholm (1967). See also Williamson (2013 [1990]: 126-143) and Salmon's (1993) reply.

4 Arguments from Reductionism

Over the last two sections I showed how and where wild extrinsicness in personal identity will arise. I argued that goodness and badness cannot plausibly be wildly extrinsic like that. So, goodness and badness cannot depend on personal identity. I think the argument works more generally to show that personal identity is not what matters. That, I take it, is the best version of Parfit's argument from extrinsicness.

I will now try to find the best version of Parfit's argument from reductionism. By "reductionism" Parfit means the claim that "a person's existence just consists in the existence of a body, and the occurrence of a series of thoughts, experiences, and other mental and physical events" (1995: 16).[36] How do we get from there to a claim about what matters?[37]

Parfit suggests multiple routes. I will show what they are, identify their pros and cons, and show a way that avoids all the cons and keeps all the pros.

4.1 Parfit's Argument from Indeterminacy

The first route starts from the idea that, given reductionism, persons are relevantly like heaps of sand: supervenient macro entities with no privileged micro basis. So, just as there sometimes is indeterminacy about how many grains it takes to make a heap, there is indeterminacy about how much mental and

[36] See also Parfit (1987: 209-217).

[37] In the case of his argument from reductionism, unlike in the case of his argument from extrinsicness, Parfit seems to draw conclusions not just about justified self-interested concern but directly about other factors of rational or moral importance.

physical continuity it takes for a single person to persist through time. As an example, consider

> **The Combined Spectrum**. Derek can undergo one of the following 100 possible operations. In Operation 1, 1% of Derek's body and brain is destroyed and replaced by the corresponding parts of Greta Garbo's body as she was at age 30. The resulting person has almost all of Derek's memories, intentions, and so on, but also a little bit of Garbo's. In Operation 2, 2% of Derek's body and brain are so replaced. And so on. In Operation 100, 100% of Derek's body and brain is destroyed and replaced by Garbo's body and brain. The resulting person has none of Derek's memories, intentions, and so on, but all of Garbo's.[38]

It seems that persons and heaps can be indeterminate in the same sorts of ways. Is that a problem if identity is what matters? That depends on how that sort of indeterminacy is to be understood. I take it that Parfit understood it as having the following two key features.[39]

> **Nonclassicism**. In indeterminate cases, whether one survives or not aren't different possibilities each of which might happen, claims about whether one survives are neither true nor false, and there is no fact of the matter about what will happen.
>
> **Resolution-by-Stipulation**. Our concept of personhood isn't yet defined in indeterminate cases. It is possible to extend it to such cases by arbitrary stipulation.

Here's a representative passage about indeterminacy in clubs:

---

[38] This is a version of Parfit's (1987: 236-7) case.

[39] Parfit's picture seems to be shared by Swinburne (1974) and Johnston (1992).

> "we can ask, 'Is this the very same club, or is it merely another club, that is exactly similar?' But these are not here two different possibilities, one of which must be true. When an empty question has no answer, we can decide to *give* it an answer. We could decide to call the later club the same as the original club" (1987: 214).

I agree with Parfit that this sort of indeterminacy in personal identity would be a problem for those who think personal identity is what matters. To see why, consider the following two cases.

> **Old Operation 50**. Derek goes through Operation 50, where 50% of his body and brain is replaced by Garbo's. The post-op person has some of Derek's character but also some of Garbo's.
>
> **New Operation 50**. Same as Old Operation 50, except that before the operation Derek undergoes a painful procedure so that the post-op person feels much better.

If nonclassicism is right, whether Derek survives to benefit from his sacrifice or dies with no payback aren't different possibilities. But suppose that there is extra badness when people get no payback for their sacrifices. Then whether that extra badness is present or absent aren't different possibilities either. That is implausible. Similarly, since it is neither true nor false that Derek survives to benefit from his sacrifice, it is also neither true nor false that that extra badness is present. Since knowledge requires truth, we can know everything there is to know about Old and New Operation 50 without knowing whether Derek's sacrifice was good or bad. That is also hard to believe.[40]

---

[40] To use Wasserman's (2012) image: it is implausible that omniscient and cooperative God would ever simply shrug his shoulders when asked about good and bad.

And if indeterminacy in personal identity could be resolved by arbitrary stipulation, then so could corresponding indeterminacy in goodness and badness. That is implausible, too. As Chang (2002) puts it in a different context:

> "the resolution of a borderline case lacks what we might call "resolutional remainder": given all the admissible ways in which the case might be resolved, there is no further question as to how resolution should proceed — any admissible resolution will do" (684).

And, she adds, hard cases where good and bad are at stake always have a resolutional remainder.

Sadly, Parfit's picture of indeterminacy is incoherent. Neither heaps nor persons can have that sort of indeterminacy. And there is no reason to think that the sort of indeterminacy that personal identity can have in cases like the Combined Spectrum should be troubling for those who think personal identity is what matters.[41] Luckily, I think we can find an improved version of Parfit's argument against the importance of personal identity. I will explain what it is in section 4.3 below.

So, why is Parfit's picture incoherent? Firstly, the problem with nonclassicism is, as Wright (2001) puts it, that the rhetoric of no fact of the matter "is simply inconsistent with the most basic constitutive principles concerning truth and negation" (87).

Take, for example, Parfit's claim that Derek's survival and non-survival aren't different possibilities in the middle cases of the combined Spectrum. This is

---

[41] Schoenfeld (2016) and Dougherty (2014) argue there might still be a problem with indeterminacy in what matters, even if we reject theses like nonclassicism and resolution-by-stipulation. I don't think they are right, as I argue in other unpublished work.

naturally read as saying that it is not the case that: either Derek survives or he doesn't. But it follows from basic logic that not-(A or B) is equivalent to not-A and not-B. So, Parfit's claim is equivalent to: Derek doesn't survive and it is not the case that he doesn't survive. Contradiction![42]

Can Parfit at least keep his thesis that claims about survival in the middle of the Combined Spectrum are neither true nor false? Hardly. That would mean rejecting a very basic principle about truth and falsity, namely,

Tarski's Equivalence. It is true that p iff p. It is false that p iff not-p.[43]

Put differently: to say that it is true that p is no more than to say p, and to say that it is false that p is no more than to say not-p.

To see why this poses a problem for Parfit, suppose that it is neither true nor false that Derek survives. By Tarski's Equivalence, if it isn't true that Derek survives, then Derek doesn't survive. And, again by Tarski's Equivalence, if it isn't false that Derek survives, then it is not the case that Derek doesn't survive. So, it follows from our supposition that Derek doesn't survive and that it is not the case that he doesn't survive. Contradiction![44]

This shows we should give up on nonclassicism. Even in the middle of the Combined Spectrum Derek either survives or not, and it is either true or false that he does.[45]

---

[42] This sort of argument is well-known. See Williamson (1992, 1994: 185-215).

[43] Issues about the liar paradox are orthogonal here.

[44] This sort of argument is also well-known. See Williamson (1992, 1994: 185-215), Horwich (1990: 76-7), Barnett (2009).

[45] Of course, we might reject the basic logic I assumed. But I think it is best for ethicists to stick with the same classical logic that working mathematicians and physicists are happy to use. To paraphrase Williamson (1997: 215): humans are better at logic than at ethics.

Similarly, I think we should dismiss the idea that indeterminacy in heaps, clubs or persons can be resolved by arbitrary stipulation. After all, stipulation is change of topic. As Williamson (1994) puts it:

> "We cannot stipulate truth-values for [claims made before a given stipulation takes effect] any more than we can stipulate birth-dates for the emperors of Rome. For certain purposes we may choose to treat them as true, or as false, but that does not make them true, or make them false. Stipulations don't answer old questions; they enable us to ask new and sometimes better ones" (214).[46]

Indeed, it is incoherent to think otherwise. Here is a simple argument.[47] Suppose that it is indeterminate whether 50% of shared body and brain is enough for Derek's survival in one of the spectrum cases. And suppose we can resolve this indeterminacy by arbitrary stipulation.

Say I stipulate that "is enough brain and body for Derek's survival" applies to 50%, while you stipulate that "is enough brain and body for Derek's survival" doesn't apply to 50%. If stipulation isn't a change of topic, then you and I can mean the same thing by "is enough brain and body for Derek's survival", presumably, having the property of being enough brain and body for Derek's survival.

But note that "is enough brain and body for Derek's survival" applies to 50% if, and only if, 50% has the property that "is enough brain and body for Derek's

[46] See also Fodor and Lepore (1996).

[47] Johnston (1989b) puts that sort of argument on its head and infers relativism about personal identity.

survival" picks out. This is a basic truism about what it is for a predicate to apply to some object.

So, it follows from my stipulation that 50% has the property of being enough brain and body for Derek's survival, while it follows from your stipulation that 50% does not have the property of being enough brain and body for Derek's survival. But this is a contradiction!

So, the only uncontroversial sense in which we can resolve indeterminacy is by moving on to a different, perhaps more tractable topic. But it would then beg the question against those who think that personal identity is what matters to say that, if persons can be indeterminate in the same sort of way as heaps, then we are free to no longer talk about persons at all.[48]

[48] What drives Parfit to nonclassicism and, from there, to resolution-by-stipulation seems to be his mistakenly inflated picture of truth-making. For example, about the middle cases of the Combined Spectrum he asks "What could make it true that, in one case, the resulting person would be me, and in the next he would not be me?" (1987: 239), and claims that we cannot answer that question "unless we are separately existing entities, such as Cartesian Egos" (1987: 266). But why shouldn't we answer: "What makes it true that you survive in one case but not the next is that, in the former, there is enough mental and physical continuity between you and the post-op person, but not in the latter"? See also Williamson's (1996) illuminating discussion of what could make indeterminate heaps heaps.

4.2 Parfit's Arguments from Below and from Merely Conceptual Facts

Parfit has another argument from reductionism that doesn't seem to depend on considerations of indeterminacy, even though he appeals to cases like the Combined Spectrum to make that argument more compelling. This is Parfit's

> **Argument from Below**. Personal identity consists in certain other facts. If one fact consists in others, then only the latter can have rational or moral importance. So, personal identity is not what matters.

Parfit adds: "What matters can only be one or more of the other facts in which personal identity consists" (1995: 29). It is tempting to read "one fact consists in another" as a claim about metaphysical reduction or supervenience. Then the argument seems to rely on a principle to the effect that derivative (non-fundamental) phenomena cannot be what matters. So understood, the argument is obviously a bad one.[49]

Garrett (1992), for example, points out that pain is morally and rationally important even if it consists in nothing more than a pattern of neuronal activity.[50] We might also worry that Parfit's argument leads straight to nihilism, as all plausible candidates for moral and rational importance seem to be metaphysically derivative.[51]

---

[49] As Parfit (2007) makes clear, the argument from below presupposes that it makes sense that X isn't what matters but Y is, even if X and Y always (perhaps even necessarily) go together, as X reduces to Y.

[50] Similar comments are made by Shoemaker (1985), Sosa (1990), Garrett (1998: 1998: 83-94), Johnston (1992). The label "argument from below" first appears in Johnston (1997).

[51] See Johnston (1992, 1997).

In his (1995) Parfit disowns this reading of the argument from below, suggesting that when personal identity consists in other facts, this is instead "a closer and partly conceptual relation":

> "Claims about personal identity may not mean the same as claims about physical and/or psychological continuity. But, if we knew the facts about these continuities, and understood the concept of a person, we would thereby know, or would be able to work out, the facts about persons" (1995: 33).

Hence, he adds, "questions about personal identity should be taken to be questions, not about reality, but only about our language". They are, as he often puts it, merely conceptual (at least in relation to the underlying facts of mental/physical continuity). We can summarize this as Parfit's

> **Argument from Merely Conceptual Facts**. Facts about personal identity are merely conceptual. Facts about what matters aren't merely conceptual. So, personal identity is not what matters.

Sadly, Parfit's argument is untenable even on that reading. To see this, we have to probe Parfit's idea of merely conceptual facts a bit more.

The thought seems to be that, even though facts of personal identity depend on some lower-level facts, the relevant bridge principles linking the two are merely conceptual rather than worldly. So, if we agree on lower-level phenomena but disagree on personal identity, we must be disagreeing about something merely conceptual. Nothing of moral relevance can depend on that sort of disagreement. Bridge principles from physics to biology, say, aren't like that.

Parfit's layered picture of reality, with some bridge principles more worthy than others, seems to be widely shared. For example, Sidelle (2007) says that insofar as a bridge principle

> "isn't a causal claim (which can be easily discerned), or claim of some contingent connection between the base facts and some further fact (in which case, these will not really be base facts (…)), these sorts of claims can only be made true by linguistic facts, and where the language is indeterminate, there will be no such facts" (110).[52]

But that sort of picture faces some serious problems. First: we might doubt whether we can meaningfully distinguish conceptual from other truths.[53]

Second: we might doubt whether putting truths about personal identity on the conceptual side amounts to *deflating* personal identity rather than *inflating* conceptual truth. If personal identity is somehow a merely conceptual matter, that might just show that facts of moral and rational importance can be merely conceptual after all.[54]

Third: we might worry that the argument from merely conceptual facts might again overgeneralize, much like the argument from below. If we agree with Parfit about the merely conceptual character of bridge principles linking personal identity to the base, we might also agree with Chalmers (1996) that biological, architectural, astronomical, chemical, economic and sociological facts "are not the sort of thing that can float free of their physical underpinnings even as a conceptual possibility" (73).

---

[52] Similar views are endorsed by Chalmers (1996: 71-89, 2011: 535-8), Horgan (1997), Heller (2008). See also Hawthorne (2009) for some relevant critical discussion.

[53] Doubts about conceptual truth were forcefully raised by Quine (1951) and, more recently, by Williamson (2007: 48-133).

[54] Perhaps Parfit thinks that we can arbitrarily change the content of a conceptual bridge principle. But that idea leads to the same problems as resolution-by-stipulation from the last section.

Fourth and last: we might doubt whether bridge principles for personal identity are conceptual in the first place. We might instead agree with Schaffer's (2017) general claim that "in all concrete transitions from more to less fundamental, the dependence functions involved provide substantive information" (10). After all, for all microphysics tells us, there might be no such macro entities as tables and persons at all. It seems to be a substantive discovery that there are bridge principles which ground the existence and character of macro entities.[55]

[55] Johnston's (2010: 306-316) reply to Parfit is instead that facts about personal identity aren't merely conceptual, since the nonexistence of souls is an empirical, rather than a conceptual, discovery. But that reply is slightly off-target: for all Johnston said, Parfit might still be right *on the assumption of a soulless reductionist view.*

4.3 My Argument from Nonsubstantiveness

There is another sort of indeterminacy that might be present in personal identity. As I will argue, it is naturally seen as depriving personal identity of rational and moral importance, without drawing on a dubious picture of indeterminacy or metaontology dismissive of higher-level phenomena. It is modelled on the sort of indeterminacy that Field (1973) claims to find in Newtonian physics.

Field's first step is to note the centrality of the following two principles in Newtonian physics.

(1) Momentum is mass times velocity.

(2) Mass of a body doesn't vary across frames of reference.

The problem is that it follows from Einstein's theory of special relativity that nothing in the world satisfies them both. Instead, there are two distinct quantities: *relativistic mass* and *proper* (or rest) *mass*. The former satisfies the first principle, the latter satisfies the second. Special relativity does not privilege one over the other. If it did, we should perhaps say that mass has really been the more privileged quantity all along. And special relativity does not recognize a "disjunctive" quantity which combines the relevant features of relativistic and proper mass. But we don't want to say that there really is no such thing as mass after all. For example, if Newton said "the Sun has more mass than the Earth", he would have said something true.

So what is mass, then? A natural Fieldian proposal is that that's indeterminate: it is determinate that mass is either relativistic mass or proper mass, but indeterminate which. This sort of indeterminacy is quite different from that

present in spectrum cases. And it is also fully compatible with classical logic, while leaving no room for resolution by stipulation.[56]

Still, it is natural to think there is nothing important left to settle about once we have settled that nothing can be mass as Newton saw it, we agree on which properties can do parts of the job Newton wanted mass to do, and how well each of them carves the world at its joints.

A bit more abstractly, we can say that, first, nothing satisfies the mass role, understood as the description of everything that mass was supposed to do in Newton's theory.[57] Second, the mass role has imperfect satisfiers that are just as good as each other and just as natural, on some scale of naturalness ranging from gerrymandered properties and relations, such as grueness, to perfectly natural ones, such as electronhood.[58] And, so, lastly, we might say that it is *nonsubstantive* whether some object a has property F if, and only if, there are equally good and equally natural (possibly imperfect) satisfiers of the F role, say, G and H, and a is G but not H. Relations of greater arity are handled similarly.[59]

But now note that this sort of indeterminacy (and nonsubstantiveness) might also be present in cases of fission and superlongevity. I will focus on the former because of their greater familiarity.

---

[56] "Fieldian" doesn't mean "Field's". I ignore some irrelevant complications of Field's own proposal.

[57] Compare Lewis (1972). Also relevant is Lewis's discussion of imperfect and multiple realizations of a theoretical role in his (1997, 1999: 291-324).

[58] On naturalness in this context, see Lewis (1983), and Dorr and Hawthorne's (2013) comprehensive survey.

[59] This is roughly Sider's (2011: 44-66) account of nonsubstantiveness, except that Sider's account is needlessly, I think, cast in metalanguage rather than object language.

First, our usual ways of identifying people across time and at a time cannot both be right in cases like that. And they all seem just as central to our best theory of personal identity in the ordinary run of cases. So, nothing in the world satisfies the person role perfectly. Still, it is natural to think there are multiple equally good imperfect satisfiers. These are the multiple relations that different accounts of personal identity take to be personal identity itself: the relation of closest continuation, the relation of belonging to some maximal psychologically-interrelated aggregate of person-stages, the relation of psychological continuation, and so on. And it seems that none of these relations is more natural than others.[60]

---

[60] Hawley (2005) thinks otherwise. She argues that termination views of fission, such as Nozick's (1981: 29-70) closest-continuer theory, lead to implausible noncausal counterfactual correlations. This makes their candidate relations especially unnatural. I disagree: I think all views of fission face similar choices here. Start with termination views. Lefty is obviously Derek in Single Transplant but not in Double Transplant. So, assuming that distinctness is necessary, Lefty in the former case isn't Lefty in the latter case. Hence, we can truly say to Lefty in Double Transplant: "Lucky you! You wouldn't have existed if the other transplant failed to take". Compare Noonan (1985, 2019: 136-7). But a similar problem arises for cohabitation views. Imagine a case like Double Transplant except where neither transplant succeeds. There is just one person in that case. If it is Lefty from Double Transplant, then, by necessity of distinctness, it cannot also be Righty from Double Transplant, since the two are different people in Double Transplant. So, we can then truly say to Righty before the operation in Double Transplant: "Lucky you! If that future operation didn't succeed, you wouldn't have been born at all". That is another implausible non-causal counterfactual correlation! And if it isn't Lefty who is present in a case where neither transplant succeeds, we can say the same thing to Lefty in Double Transplant. Now consider multilocation views and recall cases of Double Graft and Single Graft. Lefty and Righty are the same person in Double Graft, but not in Single Graft. By necessity of distinctness, Righty in the former case isn't Righty in the latter case. So, we can truly say to Righty before the operation in Double Graft: "Lucky you! If that future operation didn't succeed, you wouldn't have

Since there are persisting people, we can conclude, like with mass, that it is determinate that personal identity is one of the imperfect satisfiers of the person role but indeterminate which one.

Johnston (1992) argues that fission cases are indeterminate in precisely that way:

> "When a case necessarily violates some principle relatively central to our conception of persons and their identity over time, the concepts of a person and of being the same person over time may not determinately apply in that case, so that there may be no simple fact about personal identity in that case" (603).[61]

I think it follows, like with mass, that personal identity isn't a substantive matter in the relevant cases.[62]

That would be a problem for those who think that personal identity is what matters. Take, for example, the claim that it is bad when people suffer uncompensated burdens. And consider the following two cases.

---

been born at all". Compare Dainton (2008: 403). Another implausible non-causal counterfactual correlation! So, all views of fission have to choose between necessity of distinctness and correlations of that sort.

[61] Johnston (1989a, 1992, 1997) actually thinks that positing indeterminacy in cases like fission helps his overall case against Parfit. But that is because, I think, he sees that indeterminacy in nonclassical, stipulation-friendly terms. See Garrett (2004) for more on Johnston.

[62] We might say that fission is a case of *overdetermination*: we put too many demands on the concept of personal identity. By contrast, Sider (2001) discusses cases of *underdetermination* where our intuitions about personal identity run dry. Both sorts of cases might be thought to lead to problematic indeterminacy. But in underdetermination cases there is always the possibility that our intuitions run dry simply because we haven't thought about the matter enough. Better to focus on overdetermination cases where it is easier to see that we are demanding too much. So, my approach here is more similar to Eklund (2002) than to Sider (2001).

Sacrifice. Same as Double Transplant except that, before the operation, Derek undergoes a sacrifice for the sake of Lefty's later benefit.

No Sacrifice. Same as Double Transplant. Derek undergoes no sacrifice for the sake of Lefty's later benefit.

If it isn't a substantive matter whether Derek is Lefty in Double Transplant, then it isn't a substantive matter whether there is any extra badness in Sacrifice that is missing in No Sacrifice. So, two parties could disagree on the presence of that badness while agreeing on what the world is fundamentally like, on how to best carve it at its joints, and on the relative pros and cons of the different possible theories of personal identity, at least in the ordinary run of cases. That is implausible.[63]

So, at last, I think we found a successful version of Parfit's argument from reductionism:

Argument from Nonsubstantiveness. In cases of fission personal identity isn't a substantive matter. What matters is never like that. So, personal identity is not what matters.

To paraphrase one of Parfit's (1987: 277) comments about the Combined Spectrum:

> There is sometimes a real difference between some future person's being me, and his being someone else. But there is no such real difference in the cases of fission and fusion. What could the difference be?

[63] We might even worry that the two parties have incompatible but equally good moral concepts. Eklund (2012) discusses similar problems, albeit in the context of metaethics.

In ordinary cases, when we are told that two people are really the same, we learn that some theoretically important role is instantiated, or that some especially natural relation is present, or both. But that isn't so in fission cases where personal identity is theoretically overloaded and splits into multiple successor concepts. Each of them keeps the original concept's theoretical role and place in the hierarchy of naturalness. So, we see that it isn't personal identity as such that matters.

It is also worth noting that my argument doesn't imply that personal identity is nonsubstantive in the middle cases of the Combined Spectrum. This is because these cases don't theoretically overload personal identity. So, in these cases, there is the best and most natural satisfier of the person role. In the middle cases of the Combined Spectrum there is indeterminacy in whether Derek survives the operation, but it is a different sort of indeterminacy than in the case of mass or personal identity in fission.[64] Not all indeterminacy is nonsubstantiveness.

[64] It implies, however, that there can be indeterminacy in which relation best and most naturally satisfiers the person role. So, there will be indeterminacy in best satisfaction or in naturalness or in both. See Sud (2018) for more on indeterminate naturalness.

5 Values without persons

I defended two arguments for the claim that personal identity is not what matters. The first is a version of Parfit's

> **Argument from Extrinsicness**. Personal identity is a wildly extrinsic matter. What matters is never like that. So, personal identity is not what matters.

I showed how it works in fission and fusion cases, but also in cases of superlongevity. My second argument is a version of Parfit's argument from reductionism, namely,

> **Argument from Nonsubstantiveness**. In cases of fission personal identity isn't a substantive matter. What matters is never like that. So, personal identity is not what matters.

In both cases I focussed on how personal identity matters for value, although I think my arguments generalize to other factors of moral importance. So, what does the resulting value theory look like?

The immediate upshot of my arguments is that no goodness or badness can depend on how person-stages are packaged into people's lives. But this doesn't mean that no goodness or badness can depend on how person-stages relate to each other in terms of mental and physical continuity. So, for example, for all I argued, we might calculate the value of the world by adding up the goods and bads of person-stages but then add, on top of that, the goods or bads of certain patterns spanning multiple person-stages, such as narratives of failure or redemption.[65]

---

[65] These sorts of values are defended by Slote (1982) and Velleman (1991).

So, my arguments don't lead straight to

> **Complete Utilitarianism**. The value of the world is the total value of person-stages in that world.[66]

There is still room for other values such as value of achievement, equality among person-stages, their average value, and so on.

But there is one noteworthy implication of my arguments that does lend some credence to complete utilitarianism. Consider the following three worlds:

> **World *A***. 10 billion people live for 100 good years each and then die.
>
> **World *B***. 365 trillion people live for one good day each and then die.
>
> **World *C***. Same as world B except that everyone is slightly happier.

Let's also assume that people in world A lead fairly psychologically disunified lives while still remaining the same people throughout. Perhaps they get brainwashed every year. Then, since no further goodness or badness can depend on how person-stages are packaged into lives, it is plausible to think that worlds A and B are equally good. After all, the only major difference between them is that the former packs 365 trillion one-year-long person-stages into 10 billion fairly psychologically disunified lives. And since world C is uncontroversially better than world B, it then follows, by transitivity, that world C is better than world A.

But since people's lives in world C are arguably barely worth living (because they are so short!), while those in A are very well worth living, this is an instance of

[66] See Broome (2004: 110).

> **The Repugnant Conclusion**. For any outcome in which many people exist, all with very good lives, there is a better outcome in which many more different people exist, all with lives barely worth living.[67]

We can no longer condemn all instances of the repugnant conclusion as repugnant. Since their alleged repugnance has been a major stumbling block to wider acceptance of complete utilitarianism, this is a boost to the credibility of that doctrine.

[67] Compare Parfit (1987: 388). So, my arguments bridge the gap between Part 3 and Part 4 of *Reasons and Persons*.

6 Conclusion

In this paper I defended two Parfitian arguments for the claim that identity is not what matters: the argument from extrinsicness and the argument from nonsubstantiveness. The former draws on cases of fission, fusion, and superlongevity. The latter draws on the specific sort of indeterminacy that might arise in fission cases. It is to be distinguished from Parfit's less successful arguments from indeterminacy, from below, and from merely conceptual facts. While my arguments support an impersonal approach to value theory, they don't lead directly to complete utilitarianism.

7 References

Bader, R. (2013). Towards a hyperintensional theory of intrinsicality. *Journal of Philosophy, 110*(10), 525-563. doi:10.5840/jphil2013110109

Bader, R. (forthcoming). The fundamental and the brute. *Philosophical Studies*, 1-22. doi:10.1007/s11098-020-01486-z

Barnett, D. (2009). Is vagueness sui generis? *Australasian Journal of Philosophy, 87*(1), 5-34. doi:10.1080/00048400802237376

Beardsley, M. C. (1965). Intrinsic value. *Philosophy and Phenomenological Research, 26*(1), 1-17. doi:10.2307/2105465

Briggs, R., & Nolan, D. (2015). Utility monsters for the fission age. *Pacific Philosophical Quarterly, 96*(3), 392-407. doi:10.1111/papq.12079

Broome, J. (1991a). Utilitarian metaphysics? In J. Elster, & J. Roemer (Eds.), *Interpersonal comparison of well-being* (pp. 70–97). Cambridge: Cambridge University Press.

Broome, J. (1991b). *Weighing goods: Equality, uncertainty and time*. Oxford: Basil Blackwell.

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

Brueckner, A. (1993). Parfit on what matters in survival. *Philosophical Studies, 70*(1), 1-22. doi:10.1007/bf00989659

Campbell, T. *Personal identity and aggregation.* Unpublished manuscript. https://www.academia.edu/8854258/Personal_Identity_and_Aggregation

Chalmers, D. (1996). *The conscious mind: In search of a fundamental theory*. New York; Oxford: Oxford University Press.

Chalmers, D. (2011). Verbal disputes. *The Philosophical Review, 120*(4), 515-566. doi:10.1215/00318108-1334478

Chang, R. (2002). The possibility of parity. *Ethics, 112*(4), 659-688. doi:10.1086/339673

Chisholm, R. M. (1967). Identity through possible worlds: Some questions. *Noûs, 1*(1), 1-8. doi:10.2307/2214708

Dainton, B. (1992). Time and division. *Ratio, 5*(2), 102-128.

Dainton, B. (2008). *The phenomenal self*. Oxford University Press.

Dorr, C., & Hawthorne, J. (2013). Naturalness. In K. Bennett, & D. Zimmerman (Eds.), *Oxford studies in metaphysics, volume 8* (pp. 3-77) Oxford University Press.

Dougherty, T. (2014). Vague value. *Philosophy and Phenomenological Research, 89*(2), 352-372. doi:10.1111/phpr.12026

Eklund, M. (2002). Personal identity and conceptual incoherence. *Noûs, 36*(3), 465-485. doi:10.1111/1468-0068.00380

Eklund, M. (2012). Alternative normative concepts. *Analytic Philosophy, 53*(2), 139-157. doi:10.1111/j.2153-960X.2012.00559.x

Field, H. (1973). Theory change and the indeterminacy of reference. *The Journal of Philosophy, 70*(14), 462-481. doi:10.2307/2025110

Field, H. (2000). Indeterminacy, degree of belief, and excluded middle. *Noûs, 34*(1), 1-30. doi:10.1111/0029-4624.00200

Fodor, J., & Lepore, E. (1996). What cannot be evaluated cannot be evaluated and it cannot be supervalued either. *The Journal of Philosophy, 93*(10), 516-535. doi:10.2307/2940838

Garrett, B. (1990). Personal identity and extrinsicness. *Philosophical Studies, 59*(2), 177-194. doi:10.1007/BF00368205

Garrett, B. (1992). Persons and values. *Philosophical Quarterly, 42*(168), 337-344.

Garrett, B. (1998). *Personal identity and self-consciousness*. London: Routledge.

Garrett, B. (2004). Johnston on fission. *Sorites, 15* (December), 87-93.

Hawley, K. (2005). Fission, fusion and intrinsic facts. *Philosophy and Phenomenological Research, 71*(3), 602-621.

Hawthorne, J. (2009). Superficialism in ontology. In Chalmers, D., Manley, D. & Wasserman, R. (Eds.), *Metametaphysics: New essays on the foundations of ontology* (pp. 213-230) Oxford University Press.

Heller, M. (1987). The best candidate approach to diachronic identity. *Australasian Journal of Philosophy, 65*(4), 434-451. doi:10.1080/00048408712343071

Heller, M. (2008). The donkey problem. *Philosophical Studies, 140*(1), 83-101. doi:10.1007/s11098-008-9227-z

Horgan, T. (1997). Deep ignorance, brute supervenience, and the problem of the many. *Philosophical Issues, 8*, 229-236. doi:10.2307/1523007

Horwich, P. (1990). *Truth*. Oxford: Basil Blackwell.

Hurka, T. (1998). Two kinds of organic unity. *The Journal of Ethics, 2*(4), 299-320. doi:10.1023/A:1009795120631

Jeske, D. (1993). Persons, compensation, and utilitarianism. *The Philosophical Review, 102*(4), 541-575. doi:10.2307/2185683

Johnston, M. (1989a). Fission and the facts. *Philosophical Perspectives, 3*, 369.

Johnston, M. (1989b). Relativism and the self. In M. Krausz (Ed.), *Relativism: Interpretation and confrontation* (441-472). Notre Dame University Press.

Johnston, M. (1992). Reasons and reductionism. *Philosophical Review, 101*(3), 589.

Johnston, M. (1997). Human concerns without superlative selves. In J. Dancy (Ed.), *Reading parfit* (pp. 149-179) Blackwell.

Johnston, M. (2010). *Surviving death.* Princeton, N.J.; Oxford: Princeton University Press.

Johnston, M. (2017). The personite problem: Should practical reason be tabled? *Noûs, 51*(3), 617-644. doi:10.1111/nous.12159

Kagan, S. (1998). Rethinking intrinsic value. *The Journal of Ethics, 2*(4), 277-297. doi:10.1023/A:1009782403793

Korsgaard, C. M. (1983). Two distinctions in goodness. *The Philosophical Review, 92*(2), 169-195. doi:10.2307/2184924

Langford, S. (2007). How to defend the cohabitation theory. *Philosophical Quarterly, 57*(227), 212-224. doi:10.1111/j.1467-9213.2007.480.x

Langton, R., & Lewis, D. (1998). Defining 'intrinsic'. *Philosophy and Phenomenological Research, 58*(2), 333-345. doi:10.2307/2653512

Lewis, D. (1970). How to define theoretical terms. *The Journal of Philosophy, 67*(13), 427-446. doi:10.2307/2023861

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy, 50*(3), 249-258. doi:10.1080/00048407212341301

Lewis, D. (1976). Survival and identity. In A. Oksenberg Rorty (Ed.), *The identities of persons* (pp. 17-40) University of California Press.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy, 61*, 343-377.

Lewis, D. (1997). Naming the colours. *Australasian Journal of Philosophy, 75*(3), 325-342. doi:10.1080/00048409712347931

Lewis, D. (1999). Reduction of mind. *Papers in metaphysics and epistemology* (pp. 291-324). Cambridge: Cambridge University Press.

Martin, R., Barresi, J., & Giovannelli, A. (1998). Fission examples in the eighteenth and early nineteenth century personal identity debate. *History of Philosophy Quarterly, 15*(3), 323-348.

Noonan, H. W. (1985a). The closest continuer theory of identity. *Inquiry, 28*(1-4), 195-229. doi:10.1080/00201748508602052

Noonan, H. W. (1985b). Wiggins, artefact identity and 'best candidate' theories. *Analysis, 45*(1), 4-8. doi:10.2307/3327395

Noonan, H. W. (2019). *Personal identity* (Third ed.). Abingdon, Oxon; New York, NY: Routledge, an imprint of the Taylor & Francis Group.

Nozick, R. (1974). *Anarchy, state, and utopia* Basic Books.

Nozick, R. (1981). *Philosophical explanations* Harvard University Press.

Parfit, D. (1971). Personal identity. *The Philosophical Review, 80*(1), 3. doi:10.2307/2184309

Parfit, D. (1973). Later selves and moral principles. In A. Montefiore (Ed.), *Philosophy and personal relations* (pp. 137-169) Routledge and Kegan Paul.

Parfit, D. (1976). Lewis, perry, and what matters. In A. Oksenberg Rorty (Ed.), *The identities of persons* (pp. 91-107) University of California Press.

Parfit, D. (1982). Personal identity and rationality. *Synthese, 53*(2), 227-241. doi:10.1007/BF00484899

Parfit, D. (1987). *Reasons and persons* (2nd repr. with corrections of 1984 ed.). Oxford: Clarendon Press.

Parfit, D. (1993). The indeterminacy of identity: A reply to brueckner. *Philosophical Studies, 70*(1), 23-33. doi:10.1007/BF00989660

Parfit, D. (1995). The unimportance of identity. In H. Harris (Ed.), *Identity* (pp. 13-45) Oxford University Press.

Parfit, D. (2007). Is personal identity what matters. *Ammonius Foundation,* Retrieved from http://www.stafforini.com/docs/parfit_-_is_personal_identity_what_matters.pdf

Parfit, D. (2011). The unimportance of identity. In S. Gallagher (Ed.), *The oxford handbook of the self.* Oxford: Oxford University Press.

Perry, J. (1972). Can the self divide? *The Journal of Philosophy, 69*(16), 463-488. doi:10.2307/2025324

Prior, A. (1957). Opposite number. *The Review of Metaphysics, 11*(2), 196-201.

Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review, 60*(1), 20-43. doi:10.2307/2266637

Rabinowicz, W., & Ronnow-Rasmussen, T. (2000). A distinction in value: Intrinsic and for its own sake. *Proceedings of the Aristotelian Society, 100*(1), 33-51. doi:10.1111/1467-9264.00064

Regan, D. H. (2003). How to be a Moorean. *Ethics, 113*(3), 651-677. doi:10.1086/367589

Robinson, D. (1985). Can amoebae divide without multiplying? *Australasian Journal of Philosophy, 63*(3), 299-319. doi:10.1080/00048408512341901

Ross, J. (2014). Divided we fall. *Philosophical Perspectives, 28*(1), 222-262. doi:10.1111/phpe.12050

Salmon, N. (1993). This side of paradox. *Philosophical Topics, 21*(2), 187-197. doi:10.5840/philtopics199321219

Sattig, T. (2015). *The double lives of objects: An essay in the metaphysics of the ordinary world* Oxford University Press.

Schaffer, J. (2017). The ground between the gaps. *Philosophers' Imprint, 17*

Schoenfield, M. (2016). Moral vagueness is ontic vagueness. *Ethics, 126*(2), 257-282. doi:10.1086/683541

Shoemaker, D. W. (2002). Disintegrated persons and distributive principles. *Ratio, 15*(1), 58-79. doi:10.1111/1467-9329.00176

Shoemaker, S. (1985). Critical notice of reasons and persons by derek parfit. *Mind, 94*(375), 443-453.

Sidelle, A. (2007). The method of verbal dispute. *Philosophical Topics, 35*(1), 83-113. doi:10.5840/philtopics2007351/25

Sider, T. (1996). All the world's a stage. *Australasian Journal of Philosophy, 74*(3), 433-453. doi:10.1080/00048409612347421

Sider, T. (2001). Criteria of personal identity and the limits of conceptual analysis. *Noûs, 35*, 189-209. doi:10.1111/0029-4624.35.s15.10

Sider, T. (2011). *Writing the book of the world*. Oxford: Oxford University Press.

Slote, M. (1982). Goods and lives. *Pacific Philosophical Quarterly, 63*(4), 311-326.

Sosa, E. (1990). Surviving matters. *Noûs, 24*(2), 297-322.

Sud, R. (2018). Vague naturalness as ersatz metaphysical vagueness. In K. Bennett, & D. Zimmerman (Eds.), *Oxford studies in metaphysics, volume 11* (pp. 243–277). Oxford: Oxford University Press.

Swinburne, R. G. (1974). Personal identity. *Proceedings of the Aristotelian Society, 74*, 231-247.

Unger, P. (1990). *Identity, consciousness, and value*. New York, New York; Oxford, England: Oxford University Press.

Van Deemter, K. (2010). *Not exactly: In praise of vagueness*. Oxford: Oxford University Press.

Velleman, J. D. (1991). Well-being and time. *Pacific Philosophical Quarterly, 72*(1), 48-77. doi:10.1111/j.1468-0114.1991.tb00410.x

Wasserman, R. (2005). Humean supervenience and personal identity. *Philosophical Quarterly, 55*(221), 582-593.

Wasserman, R. (2012). Personal identity, indeterminacy and obligation. In G. Gasser, & M. Stefan (Eds.), *Personal identity: Complex or simple?* (pp. 63-81). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139028486.005

Wiggins, D. (1967). *Identity and spatio-temporal continuity* Blackwell.

Wiggins, D. (1980). *Sameness and substance* Harvard University Press.

Williams, B. (1956). Personal identity and individuation. *Proceedings of the Aristotelian Society, 57*, 229-252.

Williamson, T. (1992). Vagueness and ignorance. *Proceedings of the Aristotelian Society, Supplementary Volumes, 66*, 145-62.

Williamson, T. (1994). *Vagueness*. London: Routledge.

Williamson, T. (1996). What makes it a heap? *Erkenntnis, 44*(3), 327-339. doi:10.1007/BF00167662

Williamson, T. (1997). Imagination, stipulation and vagueness. *Philosophical Issues, 8*, 215-228. doi:10.2307/1523006

Williamson, T. (2007). *The philosophy of philosophy*. Malden, Mass.; Oxford: Blackwell Publishing.

Williamson, T. (2013). *Identity and discrimination*. (Reissued and updated from 1990 edition.) Oxford: Wiley-Blackwell.

Wright, C. (2001). On being in a quandary. relativism vagueness logical revisionism. *Mind, 110*(437), 45-98. doi:10.1093/mind/110.437.4

Chapter 2

Johnston *versus* Johnston

Abstract: Personites are like continuant people but shorter-lived. Johnston has recently argued that their existence would implode commonsense ethics and, so, they cannot exist. He concludes that broadly naturalistic accounts of our place in the world must, therefore, be wrong. I will argue that Johnston's arguments fail. To do that I propose an alternative account of intrinsicness, defend arguments from below against arguments from above, and clarify the meaning of reductionism about persons. I also show that commonsense ethics is far from unworkable even if personites are granted the same moral status as persons. I draw on Johnston's earlier exchanges with Parfit on personal identity and the place of ordinary concerns in a naturalistic world. I conclude by drawing general lessons about the ethics-metaphysics relationship and by sketching a more pressing but metaphysics-free problem that naturally arises from Johnston's discussion.[1]

Word count: 8708

[1] I would like to thank Ralf Bader, Teru Thomas, and Michal Masny.

1 Introduction

How do ordinary objects persist through time? According to *worm theory*, they do so by having temporal as well as spatial extension. For example, when you look at me now, what you see is only a part of me. I extend through time by having different temporal parts (or stages) at different times, much like I extend through space by having different spatial parts at different places. I am a *spacetime worm*. Worm theorists typically also say that, in addition to worms like you and me, there are many more gerrymandered worms such as the worm that includes all of my stages up to 2000 and yours after 2000.[1]

Persons differ from these gerrymanders in that their stages are *strongly connected* by causally mediated mental and physical similarity. So, any two stages in a worm-theoretic person are *continuous* in the sense that there is a chain of strongly connected stages between them. *Relation R* is then customarily defined as continuity and connectedness.[2]

But what about person-like worms that are shorter-lived? For example, what about a worm that consists of my stages up to 2000 but has no later stages? It is natural to think that this worm isn't a person.

Following Lewis (1983c), we can define a *person* to be a worm that is

(1) R-interrelated: all of its stages have R to all the others, and

[1] On pros and cons of worm theory, see, for example, Sider (2001a), Hawley (2001), and Magidor (2016). Lewis (1983c) was a prominent defender.

[2] I will follow Johnston, however, in ignoring the connectedness part.

(2) maximal: it is proper part of no other worm that is also R-interrelated.[3]

And, following Johnston (2016, 2017), we can define a *personite* to be a shorter-lived worm that is R-interrelated but not maximal.

The problem with personites is that they are so similar to persons that it seems impossible to grant moral status to persons while withholding it from personites.[4]

But if we grant moral status to both, we seem to face troubling overpopulation of morally considerable beings. For example, when you promise to meet me on the other side of the lake, are you not also making unkeepable promises to the masses of personites that happen to coincide with me at that time but cease to exist before I reach the other side of the lake? So, is it ever alright to make promises?[5] Johnston argues, more generally, that granting moral status to personites would make nonsense of commonsense ethics.

This is Johnston's

> **Personite Problem**. Personites have the same moral status as persons. But no plausible ethics can give them the same moral status.[6]

---

[3] Lewis's definition isn't neutral. It rules out Nozick's (1981: 29-70), Parfit's (1987: 253-266), and Dainton's (1992) accounts of what happens in cases where a person divides amoeba-like. This won't matter until section 5 below.

[4] What is moral status? According to Johnston (2017), if something has it, that warrants "(i) certain ends, such as reasonable benevolence directed toward that being and its legitimate interests, along with (ii) certain side-constraints on any other being's own pursuit of goods, constraints which rule out such things as imposing significant harms on beings with a moral status, absent compensation or consent" (621).

[5] This is a version of Johnston's (2017: 630) example.

[6] Olson (2010) discusses what is, essentially, the same problem. I will focus on Johnston's arguments, as they are more recent and more developed.

Johnston's intended conclusion is that there is something deeply wrong with worm theory and, by extension, with a naturalistic picture of the world that it exemplifies. That is potentially an important result.

Johnston has two main arguments for the claim that personites have the same moral status as persons. The first one I call

> **The Argument from No Intrinsic Difference**. There need not be any intrinsic difference between a personite and a person. Moral status is intrinsic. So, persons and personites have the same moral status.

The second argument doesn't appeal to intrinsicness of moral status but instead to claims about which nonevaluative differences are important enough to determine moral status. This is

> **The Argument from No Important Difference**. There need not be any important difference between a personite and a person. There is always an important difference between things that have moral status and the rest. So, persons and personites have the same moral status.[7]

I will argue that they both fail. To make my case I will draw on Johnston's own early work on personal identity.[8] So, I will use the early Johnston's insights to try to rebut the later Johnston's challenge. In section 2 I will show how to respond to the argument from no intrinsic difference by developing a more nuanced account of intrinsicness. In section 3 I will show how the argument from no

---

[7] As we will see, the two arguments are very different. Johnston often lumps them together, as in the following passage: "each personite is intrinsically just like some possible person or other, i.e. something which incontestably has a moral status. What then could be the basis for granting the person a moral status, but not the corresponding personite? What is so morally momentous about being maximal?" (Johnston 2017: 200).

[8] See Johnston (1989, 1992b, 1997).

important difference is just another instance of a fallacious argument that the early Johnston himself criticized. In section 4 I discuss a variant of the personite problem that Johnston thinks is especially troubling. In section 5 I argue that ethics is far from unworkable even if persons and personites have the same moral status. Section 6 shows that stage theory, an alternative to worm theory, isn't general enough to address the later Johnston's challenge. In section 7 I conclude by drawing some generalizable lessons about the ethics-metaphysics relationship and by identifying a more pressing problem that hides behind the personite problem but is quite independent of any high-level metaphysics.

2 Argument from No Intrinsic Difference

Take Tweedledee, an ordinary person. According to worm theory, Tweedledee is a spacetime worm. But now consider Tweedle, a worm that is exactly like Tweedledee, except shorter by a few years. Tweedle is also, uncontroversially, a person. They are both shown below.

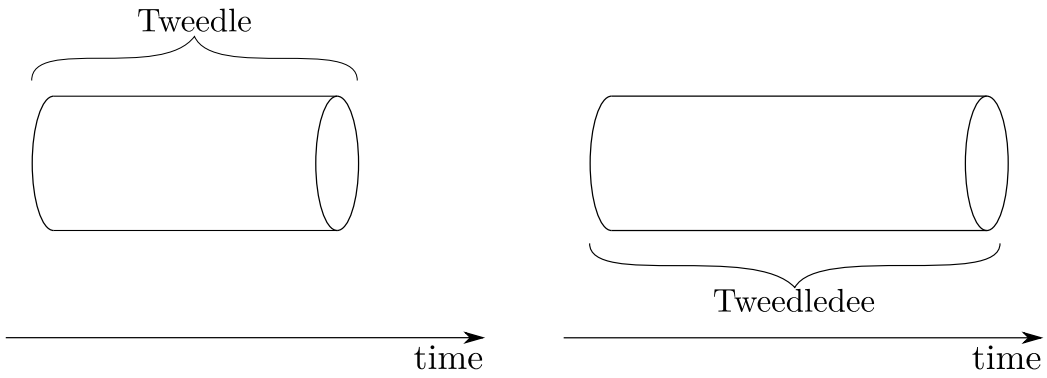


Figure 2-1. Tweedle and Tweedledee

But note that Tweedledee himself includes a shorter spacetime worm that looks very much like Tweedle. That's Tweedle*, one of Tweedledee's many personites.

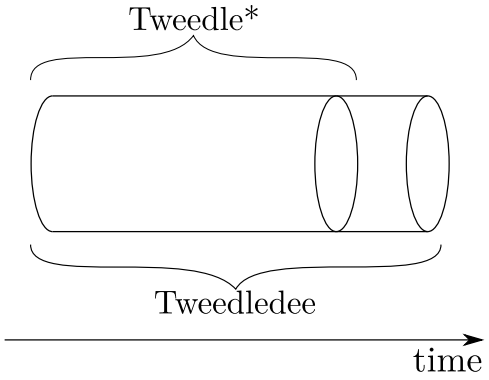


Figure 2-2. Tweedle*

Tweedle and Tweedle* look exactly the same in microphysical, chemical, biological, and even biological terms. As Johnston puts it, they are intrinsically

the same. As Lewis might have more neutrally put it, they are duplicates.[9] It is just that Tweedle*'s demise is followed by some further stages which carry on with his psychology and his body.

But moral status is plausibly intrinsic. In that respect it is "quite different from such things as being famous, being vindicated or being the best Ping-Pong player ever" (Johnston 2017: 622). So, it seems that Tweedle and Tweedle* must have the same moral status.

This is Johnston's argument from no intrinsic difference. We can usefully regiment it as follows.

(1) Persons have moral status.

(2) Moral status is intrinsic.

(3) If moral status is intrinsic, then all possible duplicates of things with moral status have moral status.

(4) There are personites.

(5) All personites are duplicates of possible persons.[10]

Therefore,

(6) All personites have moral status.[11]

If Johnston is right that no plausible ethics gives persons and personites the same moral status, we have a problem. Johnston thinks the culprit is (4) and, so, worm

---

[9] That is, there is a one-one correspondence between Tweedle's and Tweedle*'s parts (temporal and spatial) which preserves their perfectly natural relations and properties. See Lewis (1986: 61-63).

[10] This can be weakened. As Johnston (2016: 204) notes, he only needs the claim that every personite has a duplicate which has moral status.

[11] This formulation is close to that in Johnston (2016: 203-205).

theory in general. In fact, the personite problem is supposed to be an indictment of the naturalistic picture of the world, since at the centre of our commonsense ethics is "our lived sense of being *substantial* and *locally unique*, a conviction that naturalism, when understood as a complete ontology, cannot underwrite" (2017: 641). If that is right, the personite problem is serious indeed.

But, contrary to Johnston, the problem has nothing to do with worm theory or, for that matter, naturalism. We can see this by running an analogous argument about cats.

Take Tibbles the cat and Tib, a cat that is exactly like Tibbles except shorter by a tail. Tib is also, uncontroversially, a cat. They are both shown below.[12]

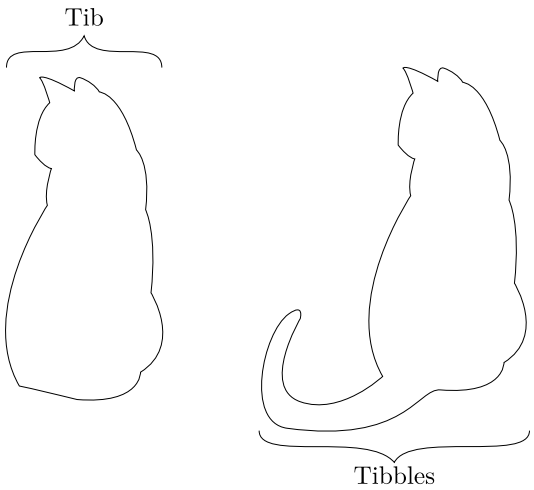


Figure 2-3. Tib and Tibbles

But note that Tibbles himself includes a smaller cat-like part that looks very much like Tib. That's Tib*.

[12] This version of Tibbles's story is told by Wiggins (1968) who credits it to Geach. A slightly different version appears in Geach (1980: 215-218)

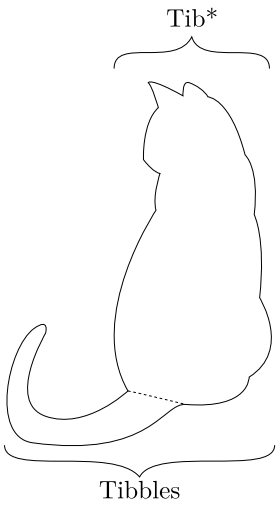


Figure 2-4. Tib*

Tib and Tib* look exactly the same in physical and chemical terms. Indeed, Johnston might say they are intrinsically the same. They are duplicates in Lewis's sense. It is just that Tib* happens to be attached to a cat tail.

We can now run an argument parallel to Johnston's own.

(1c) Cats have cathood.

(2c) Cathood is intrinsic.

(3c) If cathood is intrinsic, then all possible duplicates of cats are cats.

(4c) There are cat-like proper parts of cats.

(5c) All cat-like proper parts of cats are duplicates of possible cats.

Therefore,

(6c) All cat-like proper parts of cats have cathood.

Just like Johnston's argument leads to an implausible overpopulation of things with moral status, this argument leads to an implausible overpopulation of cats.

But the puzzle of Tib and Tibbles has nothing to do with the metaphysics of persistence. And, note, it has nothing to do with moral status. But if Johnston's

argument from no intrinsic difference works for personites, it should work just as well for cats, but also for apples, houses, tables, and so on.

And since there are many off-the-shelf resources to resist that sort of argument in the case of cats, apples, houses, and tables, we might want to use them to resist Johnston's own argument.[13]

I think the best way to go is to try to finesse our account of intrinsicness, so that the cathood of Tib carries no implications for the cathood of Tib*, and, likewise, the moral standing of Tweedle carries no implications for the moral standing of Tweedle*. This means rejecting (3) and (3c) in the two arguments above.

I think that even Johnston should see the need for such a finessed account. On the one hand, we need to mark the intuitive difference between properties like being a person, being a cat, being an apple, and so on, and obviously extrinsic properties like "being famous, being vindicated or being the best Ping-Pong player ever". On the other hand, we don't want to say that all of my personites are persons, even if they are duplicates of possible people. And we don't want to say that there are lots of cats inside my cat and lots of apples inside my apple, and so on.

The question can only be: What is this account of intrinsicness, what rationale does it have, and can it play the role we typically want intrinsicness to play?

---

[13] There are a few (somewhat overlapping) strands in the literature discussing puzzles of that kind. One is motivated directly by puzzles like that of Tib and Tibbles. See Wiggins (1968) and Burke (1994, 2003), going back to the Stoic Chrysippus. Another is motivated by more abstract concerns about supervenience and intrinsicness. See Merricks (1998, 2003), Sider (2001b, 2003), Hawley (1998, 2005), Williams (2013). The last one is motivated by the so-called problem of the many. See Unger (1980) and Lewis (1993).

As a first step, we should give up on the Lewisian

> **Duplication Account**. F is intrinsic iff for any x that is F, whenever a y is a duplicate of x, then y is F.[14]

But we can try to get a close replacement in the form of a

> **Relativized Duplication Account**. F is intrinsic iff for any x that is F, whenever a y is a duplicate of x and y is G, then y is F.[15]

We then no longer require that all duplicates of an F are F but only those which are also G. Duplication is relativized to G-ness.

I am confident that something in the ballpark of relativized duplication is the right way to defuse the puzzles of Tib/Tibbles and Tweedle/Tweedledee. I am much less confident about the details, however. I will try to defend two relativizations sufficient to block Johnston's argument from no intrinsic difference.[16]

---

[14] Compare Lewis (1983b: 355-358, 1986: 61-63).

[15] I take the idea of relativizations from Bader (ms) who ends up with a much more complicated theory than what I will suggest below. Sosa (1990: 301-305), Hawley (2005), Wasserman (2005), and Williams (2013) also search for a principled account of intrinsicness which, among other things, avoids problems like that of Tib and Tibbles. The first two are sceptical it can be found. I discuss Williams's proposal below.

[16] We might also consider Langton and Lewis's (1998) account which makes properties like cathood intrinsic without the need for relativizations. Sider (2001b) rejects it because of that. In effect, he uses cases like Tib and Tibbles to argue against the intrinsicness of cathood rather than to motivate a better account of intrinsicness.

2.1 Mereological relativization

Start with the following intuitive difference between cathood and obviously extrinsic properties such as fame. Tib's duplicates may fail to be famous due to any number of external factors. But his duplicates can only fail to be cats if extra cat parts are attached to them.[17]

To see how that difference might be relevant, recall one of Lewis's platitudes that an account of intrinsicness should aim to capture:

> "A thing has its intrinsic properties in virtue of the way that thing itself, and nothing else, is. Not so for extrinsic properties, though a thing may well have these in virtue of the way some larger whole is" (1983a: 197).

Now note that a possible world where Tib growing a tail is not a possibility where something external to Tib changes. It is a possible world where Tib himself changes. So, to check whether Tib's cathood depends on the way Tib itself is, we have to somehow set aside possibilities where extra parts are suitably attached to Tib.[18]

---

[17] At least as long as they have the right evolutionary history. I am setting aside complications to do with evolution, since they are orthogonal to Johnston's argument and irrelevant to arguably nonbiological categories such as personhood and moral status. See Sider (2001b), Hawley (2005), Williams (2013).

[18] Compare Hawley's (1998) comment that "Any object micro-indiscernible from the present, actual me is not conscious if it is suitably attached to toes, fingers, atoms and so on. But this is not dismaying, since it does not entail that *I* would not be conscious if I incorporated extra toes, fingers, atoms and so on" (842). Despite this comment she thinks that properties like consciousness and cathood are extrinsic.

How do we do that within the confines of the relativized duplication account? At a first approximation, we might try

> **The Mereological Relativization**. F is intrinsic iff for any x that is F, whenever a y is a duplicate of x and there are no other objects to which y bears the unity relation of x, then y is F.

What do I mean by "the unity relation of x"? I mean the relation that makes x the kind of thing it is. So, the unity relation of Tib might be the relation of together sustaining typical cat functions, while the unity relation of Tweedle might be relation R, continuity with some degree of connectedness.[19]

If Tib couldn't possibly change parts, then we wouldn't have to worry about that sort of relativization. In that case holding fixed facts about the arrangement and character of Tib's parts would be enough to hold fixed everything about Tib himself. So, adding the mereological relativization is a small departure from Lewis's own duplication account. Indeed, it is easy to see how someone well-versed in Lewis's metaphysics might miss the need for it. This is because Lewis's world is full of mereological aggregates which are, arguably, mereologically constant: they cannot change or gain parts.[20] Lewis's duplication account is right for them. But mereologically inconstant things also have a place in Lewis's world, and, for them, the mereological relativization is more suitable.

---

[19] Williams (2013) proposes a notion of part-intrinsicness, where F is *part-intrinsic* iff all duplicates of an F are either F or proper part of an F. Unlike my account, Williams's account rules out the intrinsicness of *holistic* and *quantitative* properties such as *being a round cat* and *being a light cat*, respectively. This is because a duplicate of a round (or light) cat might be part of a square (or heavy) cat. I think Williams's mistake is to appeal to parthood rather than to unity relations.

[20] At least setting aside the magic of counterpart theory.

But, most importantly for my purposes, the relativization blocks Johnston's argument from no intrinsic difference. I grant that Tweedle* is Tweedle's duplicate. But it is not the right duplicate to consider, since Tweedle* is followed by extra temporal parts to which it bears the relevant unity relation, relation *R*.

Here I see myself as siding with the early Johnston against the later Johnston. In his (1989a: 379-382), for example, Johnston seeks, among other things, a clear and principled formulation of the idea that personal identity is intrinsic:

> "[w]hether some process secures the survival of a given person logically depends only upon intrinsic features of the process; that is, it does not also depend on what is happening elsewhere and elsewhen" (1989: 602).[21]

Johnston notes that examples relevantly similar to Tweedle/Tweedledee and Tib/Tibbles spell problems for any simple way of spelling out that idea, precisely because they lead to unwanted and unappealing overpopulation of ordinary objects.[22]

Johnston's response is, essentially, to propose a version of the mereological relativization, so that any duplicate process considered "*is not a part of some more inclusive process which secures the survival of a person*" (1989a: 396). Similarly, in his (1992: 98), Johnston urges that qualifications dealing with maximality are "delicate matters". I think that the later Johnston gives up where the early Johnston would encourage perseverance.

---

[21] This is a version of Wiggins's (1980: 96) "only *a* and *b* rule" whose rejection is supposed to be the problem with Nozick's (1981: 29-70) closest-continuer account of personal identity.

[22] Johnston credits Mark Hinchliff with bringing these examples to his attention.

It is important to see, however, what the mereological relativization doesn't do. For one thing it doesn't imply that things of a single sort can never be nested. That would be to rule out the possibility of Pope's triple crown, mosaics made of mosaics, and Johnston's Twenty-Oners: imaginary onion-like creatures whose twenty-one layers all have separate conscious lives.[23]

The mereological relativization allows for these examples even if being a crown, being a mosaic or being a person are all intrinsic. It simply doesn't allow us to conclude that a Twenty-Oner's sixteenth layer, say, is a person simply because it has a duplicate that is clearly a person. We have to arrive at that conclusion independently. In the case of Twenty-Oners, for example, it seems enough to note that each layer has a separate mental life.

2.2 Topological relativization

Another difference between cathood and fame is that we don't have to go very far in space and time to check whether a given thing has the former, but might have to do that to check if it has the latter.[24] So, for example, it seems that whether or not Tib is a cat depends only on what's going on with Tib and some small neighbourhood of Tib's boundary.

This works even better for things which happen to be *continua*, in the sense that they are hosted by continuous regions of spacetime with parts all over that region. Continua completely fill their space.[25] If cats were continua, then to check whether

[23] On crowns see Wiggins (1980: 73), on mosaics see Sutton (2014), on Twenty-Oners see Johnston (2017: 625).

[24] Hawley (2005) and Williams (2013) also discuss that sort of idea but dismiss it too quickly.

[25] Continua in that sense are the topic of continuum mechanics.

Tib is a cat, we would only have to check what's going on with Tib and some *arbitrarily small* neighbourhood of Tib's boundary. There is no small finite region we absolutely need to check. Rather, for any given region there is a smaller one, say, half its size, that it would be sufficient to check.

A good example of a quantity that is also like that is instantaneous velocity. Recall that the velocity at a point of a particle is standardly given by the limit of change in position over change in time.

Now suppose that a particle is moving in one dimension, and its velocity at point x is v (direction doesn't matter). Does this fact depend on the interval $[x-1, x+1]$? No, since the outer parts of this interval are inessential to determining the particle's velocity at x, in that we could read off its velocity at x by considering the smaller interval $[x-0.5, x+0.5]$. The same is true of any other finite interval centered on x.

Thus, it seems that instantaneous velocity of the particle at point x doesn't depend on what happens in any *particular* region except for the point x itself. So, in that sense, it is intrinsic. But it isn't intrinsic in Lewis's duplication sense. After all, a particle moving with high velocity can have a duplicate which, at the given time, is completely stationary.[26]

This all suggests

> **The Topological Relativization**. F is intrinsic iff for any x that is F, whenever a y is a duplicate of x and some small region around y duplicates a region around x, then y is F.

---

[26] See Butterfield (2006: 723-728) whose verdict is that, in one sense, instantaneous velocity is extrinsic but, in another useful sense, intrinsic (he calls it "local"). See also Arntzenius (2000) and Smith (2003).

Of course, almost nothing is intrinsic in that sense in our world, because, as Casati and Varzi (1999) put it,

> "ordinary physical objects (i.e., objects interpreted physically as aggregates of molecules) are not strictly speaking dense and do not have boundaries of any kind (at least, not boundaries of the smooth, continuous sort countenanced by our unreflected view of the world)" (72).

The best-case scenario for ordinary objects is when facts about them depend on relatively small regions of spacetime around them. And even that might not be enough if they can be scattered across spacetime.

Importantly for us, worm-theoretic persons might well be continua, at least in the temporal dimension. That is, it might be the case that a worm-theoretic person has to have temporal parts at each time at which it exists, and can have no temporal gaps.

Then to check whether Tweedle, for example, is a worm-theoretic person, we would only need to check an arbitrarily small interval of time following its demise. So, the topological relativization might be good enough to defuse Johnston's argument from no intrinsic difference, even if it isn't good enough in general.

Indeed, the topological relativization chimes with Johnston's own gloss on intrinsic moral status:

> "Having a moral status supervenes on the mental and physical capacities and consequent operations that are present in the being's life history; what happens after (and indeed before) that life history is not relevant (except in so far as what happens before shapes what takes place in the life history)" (2016: 203).

This leaves open what happens at the boundary of a thing's history. That is the gap that the topological relativization fills: an intrinsic property need not depend on what happens *much* after or *much* before a thing's life history, but only at most on what happens (*arbitrarily*) *shortly* before or (*arbitrarily*) *shortly* after, that is to say, on what happens at the boundary.

So, in summary, we have two ways to relativize Lewis's duplication account: mereological and topological. I think they are both principled enough. They also block Johnston's argument from no intrinsic difference. But do they deserve the title "intrinsic"?

To show this, we would have to show they can play the roles for which we typically want intrinsicness. The main role that Johnston wants intrinsicness to play is in underwriting the story of humanity's moral progress in overcoming racism, sexism, and speciesism. As he asks:

> "How then may we rationally reconstruct the deeply admirable, even if shamefully belated, expansion of the protected circle, if not by way of the historically crucial appeal to intrinsic similarities?" (2016: 211).

I hope it is clear that relativized intrinsicness can underwrite that story just as well as intrinsicness in Lewis's sense.

3 Argument from No Important Difference

I conclude that Johnston's argument from no intrinsic difference is an instance of a more general problem with intrinsicness that we can solve in a principled way by means of relativizations. What about Johnston's other argument?

> **The Argument from No Important Difference**. There need not be any important difference between a personite and a person. There is always an important difference between things that have moral status and the rest. So, persons and personites have the same moral status.

So, why isn't there an important difference between persons and personites? Recall that persons are just like personites except maximal. But it is hard to think that maximality is in itself important. Since moral status cannot plausibly depend on something like that, Johnston concludes that persons and personites have to have the same moral status.

The first thing to note is that Johnston's emphasis on maximality might be an artefact of his preoccupation with Lewis's particular version of worm theory. As we saw, for Lewis:

> person $\stackrel{\text{def}}{=}$ personite + maximality.

We might wonder whether the difference between persons and personites is that important if maximality itself seems so unimportant.

But if, unlike Lewis, we think that the constitutive unity relation which unifies person stages into persons is not only symmetric and reflexive (which Lewis accepts) but also transitive (which he doesn't), we don't need to mention maximality at all.

This is because we can then use that relation to divide all person stages into equivalence classes. We have:

person $\stackrel{\text{def}}{=}$ equivalence class of person stages under the constitutive unity relation for persons.

This definition doesn't invite questions about maximality.

But, more importantly, Johnston's argument from no important difference is too close to Parfit's problematic

> **Argument from Below**. Personal identity consists in certain other facts. If one fact consists in others, then only the latter can have rational or moral importance. So, personal identity is not what matters.[27]

On any broadly naturalistic view, personal identity is bound to consist in some facts whose independent importance is hard to see. But, as Johnston (1990: 605, 1997 162-169) himself rightly emphasized, Parfit's argument seems no better than

> **The Argument from Above**. Personal identity consists in certain other facts. If one fact consists in others, then the latter can derive rational moral importance from the former. So, personal identity can be what matters.

As Sosa (1990) puts it, "it is just not clear why or in what sense the analysandum must always matter derivatively from the analysans" (321). And importance often seems to flow in the other direction. As Garrett (1992) puts it, "if we come to believe that pain just is a pattern of neuronal activity, we shall assign to such physical patterns just the importance we presently assign to pain. The importance of the analysandum will simply be transferred to its analysans" (341).

Indeed, if importance always had to flow from the analysans to the analysandum, we might worry that nothing would be of importance in a naturalistic world where

[27] It is not stated very clearly in Parfit (1987: 245-306), but see Parfit (1995: 29).

everything ultimately consists in the pattern of microphysical properties across spacetime. This was also Johnston's (1997) worry:

> "If one took the argument from below seriously, one would conclude that [any] previously valued object is not worthy of concern. Generalizing the argument, we derive nihilism" (168).[28]

The problem is that Johnston's new argument from no important difference is another argument from below. This is clear in the following passage from Johnston (2016):

> "How could maximality be crucial to having a moral status, even if it naturally enters into the four-dimensionalist's account of what it is to be a person? After all, we cannot just stipulate that only persons have a moral status" (201).

This is a mistake. We don't need to stipulate anything's moral status. It's just that, within our commonsense ethics, we think that persons are important. When we discover that person $\stackrel{\text{def}}{=}$ personite + maximality, we come to think that maximality is derivatively important.[29] To defend commonsense ethics there is no need to establish that maximality is in itself important. Why should this argument from below be any better than Parfit's?[30]

---

[28] The Parfit/Johnston exchange continues in Parfit (1995, 2007) and Johnston (2010: 305-377).

[29] We don't have to believe, as Johnston (2016: 218) alleges, that "maximality in itself is morally momentous", but only that it is derivatively morally momentous.

[30] To be fair, I think Johnston's argument from no important difference suggests an important and neglected problem, but one which is ethical rather than metaphysical. See Chapter 27 below.

4 Continuity-Variant Problem

Johnston has another argument up his sleeve. It is another variant of the personite problem, supposedly sidestepping issues of maximality entirely. To see how it goes, recall that worm-theoretic persons are united by relation R understood in terms of the holding of chains of strong connectedness. But how strong is strong enough?

To make this vivid, consider Parfit's case of

> **The Combined Spectrum**. Derek can undergo one of the following 100 possible operations. In Operation 1, 1% of Derek's body and brain is destroyed and replaced by the corresponding parts of Greta Garbo's body as she was at age 30. The resulting person has almost all of Derek's memories, intentions, and so on, but also a little bit of Garbo's. In Operation 2, 2% of Derek's body and brain are so replaced. And so on. In Operation 100, 100% of Derek's body and brain is destroyed and replaced by Garbo's body. The resulting person has none of Derek's memories, intentions, and so on, but all of Garbo's as she was at age 30.[31]

Let's say that Derek is 99% connected with the post-op person in Operation 1. But only 1% connected with the post-op person in Operation 99. And let's say that the cutoff for strong connectedness is 50%, so that anything 50% or more connected is strongly connected but nothing else is.

But, in addition to strong connectedness in that sense, there are perfectly good relations which reflect lower or higher cutoffs. For example, there is *40%-strong connectedness* that holds between stages that are 40% connected or more, there is also *60%-strong connectedness*, and so on. Any one of these relations might

[31] This is a version of Parfit's (1987: 236-7) case.

then be used to define something like relation R. For example, we can define $R_{40\%}$ in terms of the holding of chains of 40%-strong connectedness, and $R_{60\%}$ in terms of the holding of chains of 60%-strong connectedness, and so on.

And since worm theorists are happy with arbitrary aggregates of person stages, they should also be happy with aggregates of person stages that are unified by these R_x relations rather than relation R.

So, overlapping with any worm-theoretic person there will be many person-like worms which only differ in how liberal their constitutive unity relations are with respect to connectedness. Johnston calls them (*maximal*) *continuity variants*. He thinks that granting moral status to continuity-variants would cause as much trouble for ethics as granting it to personites. But how could continuity-variants and persons differ in moral status? We can put this as

> **The Continuity-Variant Argument**. The difference between persons and continuity-variants is not important. The difference between having moral status and not having moral status is always important. So, persons and continuity-variants have the same moral status.

The dialectical advantage of appealing to continuity-variants is that issues of maximality can be sidestepped on the way to Johnston's sceptical conclusions about worm theory and, more generally, naturalism.

But why should we think there is no important difference between persons and continuity-variants? Recall our simplistic assumption that 50% connectedness is enough for strong connectedness.

Then Johnston's thought seems to be that

> person $\stackrel{\text{def}}{=}$ maximality + R_x-interrelatedness + x is 50%.

But it doesn't seem in itself important whether x is set at 50 or instead at 60 or 40. We might doubt whether moral status could depend on something unimportant like that. So, persons and continuity-variants have to have the same moral status.

But this is another argument from below. As Johnston himself forcefully argued, arguments from below aren't generally good. There seems to be no special reason to accept this one. Why shouldn't we think that 50% connectedness is important derivatively even if not in itself?[32]

Perhaps Johnston has the answer. He wonders what makes it true that some particular cutoff makes for strong connectedness and, so, indirectly determines how much change a person can survive. He claims that any broadly naturalistic view of persons, such as Parfit's reductionism,

> "entails that this universal modal fact cannot arise from some fact about the essential conditions of survival of some enduring soul pellet, Cartesian ego or separately existing mental entity 'distinct from our

[32] Another problem for Johnston's argument is that, to paraphrase Sider (2002), in the Psychological Spectrum, there might be a cutoff for "strongly connected" but also for "connected enough for the persistence of something we rightly care about", provided that we rightly care about persons more than about mere aggregates of person stages. It is natural to think that these two cutoffs are the same. So, we might think that any cutoff for strong connectedness marks a morally important difference, too. It will, of course, vague where any of these cutoffs are. But this needn't change the fact they are out there. Sud (2018) makes similar comments about vague naturalness. This is, of course, not an independent argument for caring about persons, but just a defensive move that a defender of commonsense person-based ethics could make.

> brains and bodies'. It can only be a fact that arises from how we use terms like 'person' and 'numerically the same person'…" (2016: 225).[33]

And, so, since selecting the cutoff for strong connectedness is a merely linguistic matter (rather than a worldly one), nothing of importance can depend on what it is. So, continuity-variants and persons must have the same moral status.

We can put this as

> **The Argument from Merely Linguistic Difference**. The difference between persons and continuity-variants is merely linguistic rather than worldly. The difference between having moral status and not having moral status is always worldly rather than merely linguistic. So, persons and continuity-variants have the same moral status.

Is this any better than the continuity-variant argument? I don't think so. I think Johnston is wrong about what we can infer from a broadly naturalistic view of persons.

We should follow the early Johnston in allowing for *reductionism with ordinary further facts* which "means that although personal identity does not involve the persistence through time of Cartesian egos or mental substances, there are further facts of personal identity" (1997: 153), "further" in the sense that they are not identical with facts about bodily and mental continuity but merely constituted by them, much like statues might be said to be constituted by lumps of clay.

---

[33] Reductionism is the claim that "a person's existence just consists in the existence of a body, and the occurrence of a series of thoughts, experiences, and other mental and physical events" (Parfit 1995: 16). See also Johnston's useful discussion in his (1997).

That sort of reductionism is meant to contrast with Parfit's rhetoric of personal identity being *no further fact* but also with nonreductionism that posits *superlative further facts*, involving Cartesian egos, souls, and the like.

But ordinary further facts of personal identity presuppose *ordinary bridge principles* which connect facts about personal identity with facts mental and bodily continuity on which they depend.

The later Johnston's argument from merely linguistic difference seems to presuppose, however, that the only bridge principles capable of linking the two levels are either *linguistic bridge principles* (which somehow are up to us as language-users) or *superlative bridge principles* (like those that might regulate how Cartesian egos are paired with human brains). The thought seems to be that, since the latter are missing and nothing important can hang on the former, nothing important can hang on the difference between persons and continuity-variants.[34] But that is a false dilemma. We are forgetting about the possibility of ordinary bridge principles which are neither linguistic conventions nor principles of soul mechanics.[35]

---

[34] Sidelle (2007) similarly claims that insofar as a bridge principle "isn't a causal claim (which can be easily discerned), or claim of some contingent connection between the base facts and some further fact (in which case, these will not really be base facts – as, for instance, in the connection between functional or neurophysiological facts and facts about consciousness, if Chalmers [a dualist] is right), these sorts of claims can only be made true by linguistic facts, and where the language is indeterminate, there will be no such facts" (110). See also Chalmers (1996: 71-89). Horgan (1997) defends a similar view in the special case of vagueness, against Williamson (1992, 1996, 1997).

[35] Ordinary bridge principles have at least one ally in Schaffer (2017) who argues that "in all concrete transitions from more to less fundamental, the dependence functions involved provide substantive information" (10), without necessarily appealing to any superlative facts or properties.

5 Personite ethics

So far I have argued against Johnston's arguments for giving persons and personites the same moral status. But suppose I am wrong and that personites and persons do have the same moral status. Does it follow that ethics is unworkable? In the foregoing discussion, I granted this point to Johnston, but it is now time to scrutinize it.

As an example of problems that are meant to follow, consider

> **The Budapest Case**. Mark is invited to go to Budapest for three months next summer. Since he enjoys talking to the locals in their native language when abroad, he decides to take a month-long intensive course in Hungarian. As a non-Finnish foreigner he finds learning the language very difficult. When he arrives in Budapest, Mark finds it was worth the effort after all.[36]

Note that many of Mark's personites go through the trouble of learning Hungarian but cease to exist before reaching Budapest.

Intuitively, it would be wrong for Mark to coerce a large number of people to work for him for a month, just so that he can have a pleasant stay in Hungary.

But if personites have the same moral status as people, that's relevantly like what Mark is doing. Johnston concludes from cases like that that it is immoral to be prudent, that is, to make present sacrifices for the sake of larger future gains. Personites likewise appear to threaten commonsense ethics of promising, punishment, and so on.

So, is personite ethics impossible?

---

[36] This is a version of Johnston's (2016: 212, 2017: 623) example.

Johnston does admit that personites don't cause trouble for *all* ethical theories. The personite problem "might be read as an argument for [the] most primitive form of hedonistic account of the good-making features" (2017: 642) which treats persons as mere receptacles of hedonic experiences. That form of hedonism treats persons and personites equally.

But this is an overstatement. The personite problem supports, at best, not hedonism but rather *time-slice ethics* more generally, according to which nothing morally important depends on how person stages are packaged into persons. This leaves open the possibility that moral importance attaches to non-hedonic features of person stages (their knowledge of the world, for example) as well as to their mutual relationships (equality among person stages, for example). As far as I can see, deontic prohibitions, such as a prohibition on violent interference with person stages, can also stay. Time-slice ethics needn't be as crude as Johnston makes it out to be.

Still, we might wonder how much of the commonsense ethics of compensation, promising, punishment, and so on, can also be salvaged.

One appealing approach seems to be to shift from concern with persons and personites to concern with relation R, that is, to shift to what we might call *R-based ethics*.[37]

For example, instead of saying that

> (*) person stage x is bound by a promise made by person stage y if, and only if, x and y are part of the same person,

we might say that

---

[37] Both Johnston (2016: 211-215) and Olson (2010) consider this proposal. It is, of course, inspired by Parfit's (1987: 307-347) work.

(**) person stage x is bound by a promise made by person stage y if, and only if, x and y are R-related.

And instead of saying that

(†) a benefit to person stage x fully compensates a burden to person stage y if, and only if, x and y are part of the same person,

we might say that

(††) a benefit to person stage x fully compensates a burden to person stage y if, and only if, x and y are strongly R-related.

And so on. Perhaps additional adjustments can be made to accommodate the varying strength of relation R.[38]

So, why not think that much of commonsense ethics can be preserved even if persons and personites have the same moral status? Johnston's main complaint is that R-based ethics is unfairly biased against personites and in favour of persons. So, it still doesn't treat them equally. For example,

> "a person can only be compensated by experiencing or otherwise receiving the benefit in question; whereas a personite can be compensated by another being experiencing or otherwise receiving the benefit in question" (2016: 204).

But that is simply not true. To see why, we can consider cases of division and amnesia.

As an example of division, consider

[38] Making these adjustments would be necessary to treat continuity-variants on a par with persons, since it would mean that no specific degree of connectedness is privileged.

Tweedle's Division. Tweedle's body is fatally injured. Tweedle's brain is divided in half, and each half is successfully transplanted into two healthy but brainless bodies of his twin brothers. Each of the resulting people, believes that he is Tweedle, remembers living Tweedle's life, and so on. And he has a body that is very like Tweedle's.[39]

Let's call Tweedle's brothers "Lefty" and "Righty", depending on which half of Tweedle's brain they get. The story is depicted below.

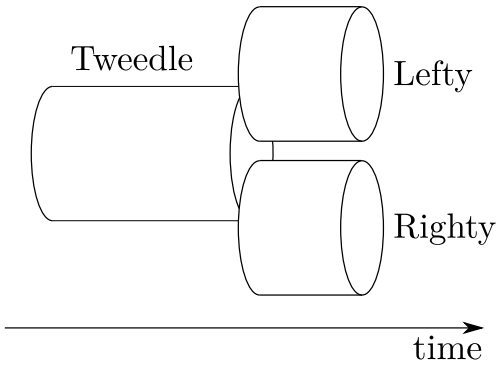


Figure 2-5. Tweedle's Division

One fairly popular take on the metaphysics of Tweedle's Division is that Tweedle ceases to exist and Lefty and Righty are two separate people different from Tweedle.[40] But according to *R*-based ethics, it won't matter that Tweedle is neither Lefty nor Righty, since Tweedle is still *R*-related to them both.

---

[39] This is a version of Parfit's (1987: 254) example.

[40] See Nozick (1981: 29-70) and Parfit (1987: 253-266). The early Johnston (1989) argued that all major accounts of what happens in cases like Tweedle's Division are on a par and, so, it is indeterminate what really happens. The later Johnston (2010: 308) thinks that the best account is that Tweedle survives as both Lefty and Righty, much like the tiger survives both in Bengal and Sumatra.

So, Tweedle might still be compensated for his burdens by a benefit that happens to Lefty. But that would be a case where a person can be compensated by another being receiving a benefit. Contrary to Johnston, this can happen to persons as well as to personites.

But perhaps Tweedle does survive in Tweedle Division. After all, wouldn't he survive if only one of the two hemispheres got transplanted?[41] If we agree, we should consider another example,

> **Tweedledee's Amnesia**. Tweedledee lives a normal life, except for frequent episodes of amnesia, mood swings and character shifts around his 40th birthday. The magnitude of these psychological changes is significant but not enough to make it the case that Tweedledee before his 40th birthday is a different person than Tweedledee afterwards.

Note that once we shift to R-based ethics it might be easier to compensate a burden to Tweedledee at age 20 by a benefit at age 30 rather than age 50, precisely because of the weakening of relation R across Tweedledee's 40th year. And that is so even if the benefit at 50 would be slightly greater than the benefit at 30.

But note that the case of Tweedledee's Amnesia also involves at least two personites: Tweedle, who expires at Tweedledee's 40th birthday, and Dee, who begins right after. This is shown in the figure below.

[41] Compare Lewis (1983c) and Dainton (1992).

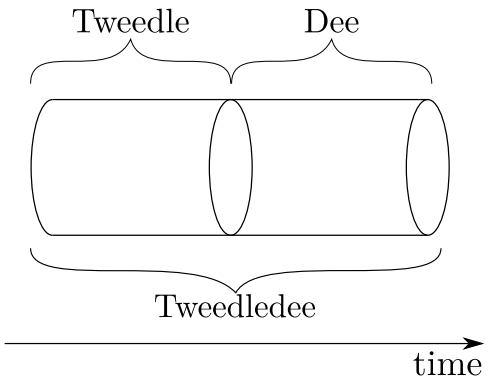


Figure 2-6. Tweedledee's Amnesia

So, when *R*-based ethics tells us to give a lesser benefit to Tweedledee at age 30 rather than to Tweedledee at age 50, it is telling us to give a greater *net* benefit to the personite Tweedle rather than to the person Tweedledee. So, this is another case where *R*-based ethics seems to favour personites rather than persons.

I conclude that *R*-based ethics has no systematic bias against personites, contrary to what Johnston claims. It sometimes favours persons, sometimes personites. Importantly, it allows us to preserve much of commonsense ethics even if personites exist and have the same moral status as persons.

6 Stage theory to the rescue?

It is worthwhile to consider another response to Johnston, one that tries to keep the duplication account of intrinsicness and commonsense ethics, but abandons worm theory in favour of *stage theory*, which says that persons are simply instantaneous (or nearly so) person stages.[42] So, according to stage theory, strictly speaking, I only exist for no more than a second. Stage theory then tries to make sense of claims like "I had breakfast in 2019" by unpacking it as: there is an instantaneous person stage in 2019, that stage had breakfast and it is a counterpart of my current stage. The relation of counterparthood is usually understood to track something like relation R that worm theorists use.

The stage theorist's response to Johnston's argument from no intrinsic difference is that *neither* personites *nor* worm-theoretic persons are *really* persons. Only person stages are persons! So, personites and worm-theoretic persons have the same moral status, namely, none. It is only person stages that have moral status. Johnston's other arguments are handled similarly.

The trouble with the stage theorist's response is that it is not general enough. This is because, as we saw with Tib and Tibbles, Johnston's argument (if it works) also works for cats, apples, houses, and people, considered as extended through space rather than time. Johnston's argument has nothing essentially to do with the metaphysics of persistence. Another problem is that stage theory shares worm theory's metaphysics. It just maps our person talk differently onto things in that metaphysics. And there are contexts where the lenses imposed by stage theory have to come off. As Sider (1996) puts it, "in certain circumstances,

[42] This response is defended by Kaiserman (2019). On stage theory itself, see Sider (2001a) and Hawley (2001).

such as when we take the timeless perspective, reference is to worms rather than stages" (448). And when it comes to morality, for example, in asking about the good and bad in the world as a whole, I think we naturally take a timeless perspective.

7 Minimalism and the Person Question

I conclude that none of Johnston's arguments work. I also argued that ethics is far from unworkable, even if I am wrong about the relative moral status of persons and personites. On all these issues I tend to side with the early Johnston against the later Johnston. What went wrong?

The early Johnston (1997) urged a kind of *minimalism*, by which he meant "the view that metaphysical pictures of justificatory undergirdings of our practices do not represent the real conditions of justification of those practices" (149-150). For example, commonsense ethics built around the notion of free will doesn't presuppose anything as exotic as libertarian free will. As a result "we can do better in holding out against various sorts of scepticism and unwarranted revision when we correctly represent ordinary practice as having given few hostages to metaphysical fortune" (150).[43]

I think that's right. But the later Johnston thinks that minimalism is powerless against personites:

> "For here the supposed ontology, when taken together with our established and *deeply admirable* ethical commitment to expand the protected circle to beings significantly like us, is at odds with something central to our ethical outlook…" (2017: 628).

I think the later Johnston is too pessimistic about what minimalism can do here. While the idea of intrinsic moral status is deeply admirable, it is also easily misinterpreted. We can and should interpret it in a way which doesn't lead to overpopulation of morally considerable beings. That is what I argued for in section 2. And once we clearly distinguish the argument from no *intrinsic* difference and

[43] Johnston's minimalism is also at work in his (1987, 1990, 1992).

the argument from no *important* difference, we can see that the latter only claims to draw on a deeply admirable ethical commitment. As a next line of defense, as we saw in section 3, commonsense ethics itself can be reinterpreted in terms of relation R, making it more robust to the threat of personites. To make a successful argument from metaphysics to ethics, we must not import unwarranted metaphysics into the ethics.[44]

Even though Johnston's arguments fail, I think they raise an important and neglected question. I call it

The Person Question. Why do persons matter?

For example, why does it matter whether a burden and its compensation, a promise and its fulfilment, a crime and its punishment, and so on, fall within the life of a single person rather within some single gerrymandered aggregate of person stages?

It seems that this problem arises independently of the metaphysics of persistence or, indeed, any high-level metaphysics. For example, even if the only things around were immaterial souls (no organisms, no psychological continuers, and so on), we might still intelligibly ask: "Why does it matter that a later gain comes to the same soul that previously suffered a burden?"

So, is there a good answer to the person question? Elsewhere Johnston (2010) himself asks "What, then, is so good or important or valuable about persons?" and answers:

> "It is because persons are thinking, reflective beings who thus experience the demand to *live* their lives, to give their lives shape

[44] I try to find a better metaphysics-to-ethics argument in my "Intrinsic concerns without extended selves" (here included as chapter 1).

> according to their idea of the good, that they deserve a kind of respect that no mere thing, however appealing, does" (269).

Whatever we think of Johnston's answer, we can see that it doesn't depend on the falsehood of worm theory. Indeed, a worm theorist might try to use it to justify a moral difference between persons and personites. It seems that we face a demand to live our *entire* lives rather than their non-maximal gerrymandered proper parts.

I am not at all sure whether Johnston's answer is right. Indeed, I am not sure how to go about answering the person question. But insofar as personites suggest a pressing problem, that is it. Unless we can solve it, we have to give up on person-based ethics and move to time-slice ethics or R-based ethics instead.

8 Conclusion

I argued in favour of the early Johnston's solutions to the later Johnston's personite problem. To do that, I introduced two relativized notions of intrinsicness, based on Lewis's duplication account. I also showed why maximality might matter derivatively, why maximal continuity-variants pose no extra ethical problems, and why R-based ethics isn't biased against personites. I also showed that stage theory won't solve Johnston's problem. Then I drew some general morals about successful arguments from metaphysics to ethics, and I sketched the real problem that reflecting on personites should make us think of.

9 References

Arntzenius, F. (2000). Are there really instantaneous velocities? *The Monist, 83*(2), 187-208.

Bader, R. (2013). Towards a hyperintensional theory of intrinsicality. *Journal of Philosophy, 110*(10), 525-563. doi:10.5840/jphil2013110109

Bader, R. (ms). "Relativised intrinsicality". Unpublished manuscript.

Burke, M. B. (1994). Dion and Theon: An essentialist solution to an ancient puzzle. *Journal of Philosophy, 91*(3), 129-139.

Burke, M. B. (2003). Is my head a person? In K. Petrus (Ed.), *On human persons* (pp. 107-125) Heusenstamm: Ontos Verlag.

Butterfield, J. (2006). Against pointillisme about mechanics. *British Journal for the Philosophy of Science, 57*(4), 709-753.

Casati, R., & Varzi, A. C. (1999). *Parts and places. the structures of spatial representation*. MIT Press.

Chalmers, D. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Dainton, B. (1992). Time and division. *Ratio, 5*(2), 102-128.

Dainton, B. (2008). *The phenomenal self*. Oxford: Oxford University Press.

Garrett, B. (1992). Persons and values. *Philosophical Quarterly, 42*(168), 337-344.

Geach, P. (1980). *Reference and generality: An examination of some medieval and modern theories* (3rd ed.). Ithaca: Cornell University Press. First edition published in 1962.

Hawley, K. (1998). Merricks on whether being conscious is intrinsic. *Mind, 107*(428), 841-843. doi:10.1093/mind/107.428.841

Hawley, K. (2001). *How things persist*. Oxford: Clarendon Press.

Hawley, K. (2005). Fission, fusion and intrinsic facts. *Philosophy and Phenomenological Research, 71*(3), 602-621.

Horgan, T. (1997). Deep ignorance, brute supervenience, and the problem of the many. *Philosophical Issues, 8*, 229-236. doi:10.2307/1523007

Johnston, M. (1987). Is there a problem about persistence? *Aristotelian Society Supplementary Volume, 61*(1), 107-135. doi:10.1093/aristoteliansupp/61.1.107

Johnston, M. (1989). Fission and the facts. *Philosophical Perspectives, 3*, 369-397.

Johnston, M. (1992a). Constitution is not identity. *Mind, 101*(401), 89-106. doi:10.2307/2254121

Johnston, M. (1992b). Reasons and reductionism. *Philosophical Review, 101*(3), 589-618. doi:10.2307/2186058

Johnston, M. (1997). Human concerns without superlative selves. In J. Dancy (Ed.), *Reading Parfit* (pp. 149-179) Blackwell.

Johnston, M. (2010). *Surviving death*. Princeton: Princeton University Press.

Johnston, M. (2016). Personites, maximality and ontological trash. *Philosophical Perspectives, 30*(1), 198-228. doi:10.1111/phpe.12085

Johnston, M. (2017). The personite problem: Should practical reason be tabled? *Noûs, 51*(3), 617-644. doi:10.1111/nous.12159

Kaiserman, A. (2019). Stage theory and the personite problem. *Analysis, 79*(2), 215-222. doi:10.1093/analys/any074

Langton, R., & Lewis, D. (1998). Defining 'intrinsic'. *Philosophy and Phenomenological Research, 58*(2), 333-345. doi:10.2307/2653512

Lewis, D. (1976). Survival and identity. In A. Oksenberg Rorty (Ed.), *The identities of persons* (pp. 17-40) University of California Press.

Lewis, D. (1983a). Extrinsic properties. *Philosophical Studies, 44*(2), 197-200. doi:10.1007/BF00354100

Lewis, D. (1983b). New work for a theory of universals. *Australasian Journal of Philosophy, 61*, 343-377.

Lewis, D. (1983c). Survival and identity, with postscripts. In D. Lewis (Ed.), *Philosophical papers, volume I* (pp. 55-77) Oxford University Press. First published in A. Oksenberg Rorty (Ed.), *The identities of persons* (pp. 17-40) University of California Press.

Lewis, D. (1993). Many, but almost one. In K. Cambell, J. Bacon & L. Reinhardt (Eds.), *Ontology, causality and mind: Essays on the philosophy of D. M. armstrong* (pp. 23-38) Cambridge University Press.

Lewis, D. K. (1986). *On the plurality of worlds* Wiley-Blackwell.

Magidor, O. (2016). Endurantism vs. perdurantism?: A debate reconsidered. *Noûs, 50*(3), 509-532. doi:10.1111/nous.12100

Merricks, T. (1998). Against the doctrine of microphysical supervenience. *Mind, 107*(425), 59-71. doi:10.1093/mind/107.425.59

Merricks, T. (2003). Maximality and consciousness. *Philosophy and Phenomenological Research, 66*(1), 150-158. doi:10.1111/j.1933-1592.2003.tb00248.x

Nozick, R. (1981). *Philosophical explanations*. Harvard University Press.

Olson, E. T. (2010). Ethics and the generous ontology. *Theoretical Medicine and Bioethics, 31*(4), 259-270. doi:10.1007/s11017-010-9148-7

Parfit, D. (1987). *Reasons and persons* (2nd repr. with corrections of 1984 ed.). Oxford: Clarendon Press.

Parfit, D. (1995). The unimportance of identity. In H. Harris (Ed.), *Identity* (pp. 13-45). Oxford: Oxford University Press.

Parfit, D. (2007). Is personal identity what matters. *Ammonius Foundation,* Retrieved from http://www.stafforini.com/docs/parfit_-_is_personal_identity_what_matters.pdf

Schaffer, J. (2017). The ground between the gaps. *Philosophers' Imprint*, volume 17.

Sidelle, A. (2007). The method of verbal dispute. *Philosophical Topics, 35*(1), 83-113. doi:10.5840/philtopics2007351/25

Sider, T. (1996). All the world's a stage. *Australasian Journal of Philosophy, 74*(3), 433-453. doi:10.1080/00048409612347421

Sider, T. (2001a). *Four dimensionalism: An ontology of persistence and time.* Oxford: Oxford University Press.

Sider, T. (2001b). Maximality and intrinsic properties. *Philosophy and Phenomenological Research, 63*(2), 357-364. doi:10.1111/j.1933-1592.2001.tb00109.x

Sider, T. (2002). Hell and vagueness. *Faith and Philosophy, 19*(1), 58-68. doi:10.5840/faithphil20021918

Sider, T. (2003). Maximality and microphysical supervenience. *Philosophy and Phenomenological Research, 66*(1), 139-149. doi:10.1111/j.1933-1592.2003.tb00247.x

Smith, S. R. (2003). Are instantaneous velocities real and really instantaneous?: An argument for the affirmative. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics, 34*(2), 261-280. doi:10.1016/s1355-2198(03)00007-8

Sosa, E. (1990). Surviving matters. *Noûs, 24*(2), 297-322. doi:10.2307/2215530

Sud, R. (2018). Vague naturalness as ersatz metaphysical vagueness. In K. Bennett, & D. Zimmerman (Eds.), *Oxford studies in metaphysics,* volume 11 (pp. 243–277). Oxford: Oxford University Press.

Sutton, C. S. (2014). Against the maximality principle. *Metaphysica, 15*(2), 381-390. doi:10.1515/mp-2014-0023

Unger, P. (1980). The problem of the many. *Midwest Studies in Philosophy, 5*(1), 411-468. doi:10.1111/j.1475-4975.1980.tb00416.x

Wasserman, R. (2005). Humean supervenience and personal identity. *Philosophical Quarterly, 55*(221), 582-593. doi:10.1111/j.0031-8094.2005.00417.x

Wiggins, D. (1968). On being in the same place at the same time. *Philosophical Review, 77*(1), 90-95. doi:10.2307/2183184

Wiggins, D. (1980). *Sameness and substance*. Harvard University Press.

Williams, J. R. (2013). Part-Intrinsicality. *Noûs, 47*(3), 431-452. doi:10.1111/j.1468-0068.2011.00837.x

Williamson, T. (1992). Vagueness and ignorance. *Proceedings of the Aristotelian Society, Supplementary Volumes, 66*, 145-62. doi:10.1093/aristoteliansupp/66.1.145

Williamson, T. (1996). What makes it a heap? *Erkenntnis, 44*(3), 327-339. doi:10.1007/BF00167662

Williamson, T. (1997). Imagination, stipulation and vagueness. *Philosophical Issues, 8*, 215-228. doi:10.2307/1523006

Chapter 3

Egalitarianism and Population Size

Abstract: According to communal egalitarianism, the value of an outcome is the result of balancing total wellbeing and the badness of inequality. According to weighted egalitarianism, it is the result of aggregating people's wellbeing while giving priority to the relatively worse-off. Unlike communal egalitarianism, weighted egalitarianism escapes the levelling-down objection and can be supported by a veil of ignorance argument. The two kinds of egalitarianism can agree in same-number cases but disagree in different-number cases. This paper presents a dilemma for weighted egalitarians. On the one hand, they have good reasons to accept limited principles of existence independence and hence a specific formula called "the geometric Gini". On the other hand, the most straightforward application of the geometric Gini in different-number cases yields counterintuitive verdicts which are at odds with egalitarian thinking. This suggests egalitarians should be communal rather than weighted egalitarians, meaning they have to give up on many promised benefits of the latter. We also see more clearly the choices we need to make if we want our value theory to reflect aversion to inequality in welfare.[1]

Word count: 9621

[1] I would like to thank Ralf Bader, Teru Thomas, Michal Masny and the audience at the graduate session of the 2019 Princeton Workshop on Population Ethics.

1 Introduction

We should distinguish between two kinds of egalitarians. *Communal egalitarians* take the goodness of an outcome to be the result of weighting two factors: the value of welfare and the disvalue of inequality.[1] Their view can naturally be expressed by the following formula:

$$\text{Value} = \text{total wellbeing} - \text{badness of inequality} \tag{1}$$

The specific behaviour of the communal-egalitarian formula will depend on how the "badness of inequality" term is defined.[2]

Weighted egalitarians, on the other hand, take the goodness of an outcome to be the result of aggregating people's wellbeing in a way which gives priority to the relatively worse-off. Their view can naturally be expressed by the formula for *the generalized Gini*:

$$\text{Value} = \alpha_1 w_1 + \alpha_2 w_2 + \cdots + \alpha_n w_n \tag{2}$$

---

[1] This kind of view is discussed in Rescher (1966: chapter 2), Weirich (1983), Persson (2001, 2008), Blackorby et al. (2005: chapter 4), Hirose (2004, 2009, 2015), and Adler (2012: chapter 5). Why "communal"? Because the formula is naturally read as attributing the badness of inequality to the community at large. The focus on the badness of *inequality* rather than the goodness of *equality* will be addressed in section 5 below.

[2] In general, we can say that the badness of inequality is zero in an outcome where everyone has the same wellbeing level; that it is positive otherwise; and that it goes down when some amount of wellbeing is transferred from a better-off person to a worse-off person without changing anyone's ranks. Compare Adler (2012: 114–115).

where the w's are wellbeing levels ordered from lowest to highest, and the α's are weights which are positive and decreasing.[3] The decreasingness of the weights ensures that this formula will prefer equal outcomes to unequal ones with the same total wellbeing. The specific sequence of weights used will determine the degree of priority given to the relatively worse-off.[4]

Should egalitarians be communal egalitarians or weighted egalitarians? Weighted egalitarianism has been argued to escape the key objection against communal egalitarianism: the levelling-down objection.[5] It has also been supported as the evaluation method that would have been selected by self-interested, risk-averse individuals behind a veil of ignorance.[6] Still, the difference between communal and weighted egalitarianism might seem obscure, as formulas (1) and (2) can generate the same ranking in same-number cases, that is, when the number of people in the outcomes compared is fixed.

We will see, however, that the difference between communal and weighted egalitarianism is like the difference between total and average utilitarianism: it only becomes clear in different-number cases, that is, when the number of people between the outcomes compared can vary.[7] In these cases, a generalized Gini

---

[3] Throughout this paper I will assume that the indexing of wellbeing levels reflects their increasing order in a given outcome. The indices therefore represent people's *ranks* in the outcome. When assigning ranks, ties are to be broken arbitrarily.

[4] The generalized Ginis were introduced by Weymark (1981). See also Donaldson & Weymark (1980). The presentation here is closest to that in d'Aspremont & Gevers (2002).

[5] See Persson (2001), Hirose (2004, 2009, 2015) Buchak (2017).

[6] See Buchak (2017).

[7] Arrhenius (2013) makes a similar claim about ways of measuring the value of inequality and equality.

formula need not produce the same ranking of outcomes as some communal-egalitarian formula.

Different-number cases also give rise to a dilemma for weighted egalitarians. On the one hand, weighted egalitarians should accept certain principles which constrain how weights in the generalized Gini formula relate to each other, both within and across population sizes. These imply that, after suitable normalizations, weighted egalitarians have to use *the geometric Gini* for comparing outcomes with the same number of people:

$$\text{Value} = w_1 + \beta w_2 + \beta^2 w_3 \ldots + \beta^{n-1} w_n \tag{3}$$

where β is some number between zero and one which effectively serves as a discount factor, dampening the wellbeing of the better-off. On the other hand, a straightforward extension of the geometric Gini to different-number cases yields some implausible verdicts which are hard to explain on egalitarian grounds.

Put differently, the dilemma for weighted egalitarians is that either they reject certain appealing principles which lead them to the geometric Gini or else they need to explain away the implausible features of the geometric Gini in different-number cases.

The first horn of the dilemma will be spelled out in section 3, and the second horn in sections 4 and 5. Section 5 will also explain why the geometric Gini cannot be motivated by the veil of ignorance. The technical material which underpins the dilemma is relegated to the appendix. The main text explains the key ideas by means of illustrative examples.

Different-number cases therefore clarify the extent and nature of the disagreement between weighted and communal egalitarians. They also suggest that weighted egalitarianism is not a well-motivated egalitarian view. This gives us reason to

think that weighted egalitarianism is not truly egalitarian, and that egalitarians have to be communal egalitarians. Hence, they cannot reap the purported benefits of weighted egalitarianism and have to face the levelling-down objection head-on.

2 Egalitarianism: weighted or communal?

This section will describe two reasons in favour of weighted egalitarianism: first, that it avoids the levelling-down objection; second, that it can be supported by a veil of ignorance argument. Then we will see that weighted egalitarianism can always be made to agree with communal egalitarianism in same-number cases. In the following sections we will see that this agreement cannot always be guaranteed in different-number cases.

2.1 Levelling-down

Some egalitarian views imply that there is something good about reducing the wellbeing of the better-off to make them level with the worse-off. For example, in the following case, outcome Y is the result of making the better-off Ann as well-off as the worse-off Bob.

	Ann	Bob
X	200	100
Y	100	100

Table 3-1

According to *the levelling-down objection*, however, Y is not better than X in any respect (let alone all-things-considered better).[8]

Which kinds of egalitarians are targeted by the objection? As Hirose (2009) explains:

> "As far as the goodness of a state of affairs is given by a function that takes the disvalue of inequality in its argument, there is always one respect with regard to which the levelling-down is better" (303).

The levelling-down objection therefore applies to communal egalitarians, but not to weighted egalitarians. The badness of inequality shows up in formula (1) but not in formula (2). So, egalitarians can avoid the levelling-down objection by becoming weighted egalitarians.[9]

This has been recognized by friends of weighted egalitarianism. For example, Hirose (2009) claims that inequality is not an object of aggregation "but a feature of an aggregative process for estimating the goodness of a state of affairs" (303). This is anticipated by Persson (2008) who claims that "the disvalue of unjust inequality [is] to be something that operates upon the 'host' value of wellbeing rather than as a separate value alongside it" (297), and echoed by Buchak (2017): "the only objects of concern are the interests of each individual, but each

[8] See Parfit (2000 [1995]; 1997)

[9] There are other possible responses to the levelling-down objection. First, bite the bullet, as do Temkin (1993), and Otsuka & Voorhoeve (2018). Second, argue that the levelling-down objection is either ill-defined or over-broad; see Broome (2002), Brown (2003) and Fleurbaey (2015). Broome (2002) also suggests the weighted-egalitarian response.

individual's interests needn't be given the same weight in the evaluation of a distribution" (624).[10]

Unlike communal egalitarians, weighted egalitarians also have a natural explanation of why levelling-down can never be better all things considered: increasing someone's wellbeing is bound to increase the overall weighted sum of wellbeing, when weights are positive.[11]

2.2 Veil of ignorance

Buchak (2017) gives another reason in favour of weighted egalitarianism: self-interested people behind a veil of ignorance would agree on the generalized Gini as a way of evaluating outcomes, provided that they are risk-averse in the right way.

We can isolate the key premise in this defence of weighted egalitarianism as

> **The Veil of Ignorance Principle**. Outcome X is at least as good as outcome Y iff the prospect of receiving the wellbeing of any person in X with equal probability is at least as good for one as the prospect of receiving

[10] The view which I call "weighted egalitarianism" has had many names: "relational prioritarianism" (Persson (2001, 2008)) "weighted egalitarianism" (Hirose (2004)), "the aggregate view" (Hirose (2015)), "relative prioritarianism" (Buchak (2017)), and "rank-discounted utilitarianism" (Asheim & Zuber (2014)). Despite the various names all these authors see weighted egalitarianism as a broadly egalitarian view.

[11] Compare Hirose (2009: 307-8). Adler (2012: 336) challenges communal egalitarians to offer such an explanation.

the wellbeing of any person in Y with equal probability, provided that X and Y have the same population size.[12]

To illustrate, consider the following two outcomes.

	People		
	Ann	Bob	Cat
Outcome X	10	20	30
Outcome Y	18	18	18

Table 3-2

As compared with outcome X, outcome Y buys perfect equality at the cost of some loss in total wellbeing. According to the veil of ignorance principle, outcome X is better than outcome Y iff, in the following table, prospect Z is better for one than prospect W.

	States of nature		
Probabilities	1/3	1/3	1/3
Prospect Z	10	20	30
Prospect W	18	18	18

Table 3-3

As compared with prospect Z, prospect W is safer since it offers 18 units of wellbeing for sure, but its expectation of wellbeing is also lower by 2 units. The comparative value of these two prospects will depend on how much safety counts against expected wellbeing.

[12] Compare Thomas (2016: 101) and Thomas et al. (2020). Buchak herself does not appeal to this principle explicitly.

Buchak (2013) defends a theory of the value of prospects for individuals, according to which safety can count for something in this comparison. More precisely, in the special case when all prospects are defined over n equally probable states of nature, Buchak's theory can be stated as follows.

$$\text{Prospective value} = \gamma_1 w_1 + \gamma_2 w_2 + \cdots + \gamma_n w_n \quad (4),$$

where the γ's are positive and decreasing weights which depend on the number of all possible states, and the w's are possible wellbeing levels one could receive, ordered from lowest to highest.[13] Combined with the veil of ignorance principle, this theory delivers weighted egalitarianism in the form of the generalized Gini formula. Buchak's theory of the individual value of prospects is popular and well-defended, even if controversial. The veil of ignorance principle is also defensible, if controversial.[14] Together they make an important case for weighted egalitarianism. By contrast, theories of the individual value of prospects which would deliver a communal-egalitarian view are less developed and more controversial.[15]

---

[13] According to Buchak's (2013) theory, the individual value of a prospect with n equally probable states can be written as:

$$\text{Prospective value} = \sum_{i=1}^{n} \left[r\left(\frac{n-(i-1)}{n}\right) - r\left(\frac{n-i}{n}\right) \right] \times w_i$$

where r is the *risk function* which reflects one's attitude to risk and n is the number of possible states of nature. Compare Buchak (2013: 56-7) and the appendix to Buchak (2017). The coefficients on wellbeing levels give the sequence of γ's in formula (4).

[14] See Harsanyi (1977), Thomas (2016), Thomas et al. (2020).

[15] See Hagen's "three moments of utility" model discussed in Sugden's (1986).

2.3 Agreement in same-number cases

Despite differences in justification, it can be shown that weighted egalitarians will agree on the ranking of outcomes with *some* communal egalitarians, at least in same-number cases.

Consider the following illustration, adapted from Hirose (2004: 82-83)[16], focussing on three-person outcomes. Take the following formula which I will simply call "the standard Gini": [17]

$$\text{Value} = \frac{5}{3}w_1 + \frac{3}{3}w_2 + \frac{1}{3}w_3 \tag{5}$$

The standard Gini is of course a generalized Gini. For an outcome with n people it uses a sequence of weights which correspond to the initial n odd numbers divided by the number of people. Simple rearrangement shows this formula to be equivalent to:

$$\text{Value} = (w_1 + w_2 + w_3) - \frac{1}{3}\big((w_2 - w_1) + (w_3 - w_2) + (w_3 - w_1)\big) \tag{6}$$

That is, total wellbeing minus the average wellbeing gap between someone and all those better-off than them. Let us call this one "the communal Gini".[18]

---

[16] Compare Blackorby & Donaldson (1978: 69), Broome (2002), and Hirose (2009, 2015).

[17] Hirose gives the general formula as:

$$\text{Value} = \frac{2n-1}{n}w_1 + \frac{2n-3}{n}w_2 + \cdots + \frac{1}{n}w_n$$

[18] Hirose gives the general formula as:

$$\text{Value} = \sum_{i=1}^{n} w_i - \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}|w_i - w_j|$$

The badness of inequality term is equal to the Gini inequality index multiplied by n times average wellbeing. See Sen (1997: 31). The equivalence between this formula and that given in the previous

Since they are mathematically equivalent, the two formulas give the same ranking of outcomes. The weighted-egalitarian view naturally expressed by (5) coincides with the communal-egalitarian view naturally expressed by (6).

This equivalence is an example of a general fact that, for every weighted-egalitarian view, there will be a communal-egalitarian view which gives the same ranking of outcomes, at least in same-number cases.[19] But the construction does not necessarily work the other way around, since it may be impossible to disperse the "badness of inequality" term into a sequence of weights which are all positive and greater for the relatively worse-off.[20]

3 Why the geometric Gini?

The key parameters of weighted egalitarianism are the weights to be used in the generalized Gini formula. How do these weights relate to each other across population sizes? And how do they relate to each other relative to a single population size?

As we will see, we can argue that the ratios of these weights have to be the same across population sizes, and that, relative to any population size, the ratio of each weight to the next should be equal to the same fixed number between zero and one. If we then set the weight of the worst-off rank to unity, we obtain the geometric Gini. It therefore looks like weighted egalitarians are in no position to

---

footnote is unobvious but mathematically elementary. It can be proved by tweaking the results in Anand (1983: 313-314).

[19] The equivalence is well-known, see Broome (2002) and Fleurbaey (2015).

[20] See Hirose (2004: 69) and Blackorby & Donaldson (1981: 70) where communal-egalitarian formulas of this kind can be found.

reject the geometric Gini in same-number cases. This is the first horn of the dilemma for weighted egalitarians.

Call a group of people *unconcerned* between outcomes X and Y if their wellbeing level in X is the same as their wellbeing level in Y. It is natural to think that the unconcerned may be ignored in comparing two outcomes. This thought is captured by

> **Independence of the Unconcerned**. If a group is unconcerned between outcomes X and Y, and between outcomes X^* and Y^*, and the remaining people are equally well-off in X as in X^*, and in Y as in Y^*, then X is at least as good as Y iff X^* is at least as good as Y^*.[21]

Weighted egalitarians cannot generally accept this principle, as can be illustrated by the following example. Take the standard Gini, as defined in formula (5) above, and consider the following four outcomes.

	Ann	Bob	Cat	Gini
X	25	20	30	≈ 68.33
Y	25	21	28	≈ 69.33
X^*	40	20	30	≈ 76.67
Y^*	40	21	28	≈ 76.33

Table 3-4

To move from X to Y is to implement a leaky transfer from Cat to Bob: two units of wellbeing are taken from Cat and one unit of wellbeing is given to Bob. The transfer is also rank-preserving in that people's relative positions are unchanged. Meanwhile Ann is unaffected by the transfer, occupying a rank between Bob and

[21] This principle of independence partly characterizes utilitarianism and prioritarianism. See Maskin (1978).

Cat. To move from X^* to Y^* is to implement the same leaky transfer from Cat to Bob, with the difference that Ann, still unaffected by the transfer, is now better-off than both Bob and Cat.

As the right-most column shows, according to the standard Gini, X is worse than Y yet X^* is better than Y^*.[22] That is, it is better to make the transfer when Ann's wellbeing is between that of Bob's and Cat's, but worse to make the same transfer when Ann is better-off than both of them. This is a violation of independence of the unconcerned.

Independence of the unconcerned fails in this example because in X and Y, Bob is two rungs below Cat on society's wellbeing ladder, while in X^* and Y^* he is merely one rung below her. While Ann is unaffected, her wellbeing level determines the relative ranks of Bob and Cat, a factor deemed relevant by the standard Gini.

But weighted egalitarians will not and need not reject independence principles *in general*.[23] One principle they will accept is

> **Independence of the (Unconcerned) Worst-Off**. If the worst-off person is unconcerned between X and Y, and between X^* and Y^*, and the remaining people are equally well-off in X as in X^*, and in Y as in Y^*, then X is at least as good as Y iff X^* is at least as good as Y^*.

That is, a comparison of two outcomes which leaves the worst-off unaffected does not depend on the wellbeing level of the worst-off.

---

[22] Calculations. Value of $X = \frac{5}{3} \times 20 + \frac{3}{3} \times 25 + \frac{1}{3} \times 30 \approx 68.33$; value of $Y = \frac{5}{3} \times 21 + \frac{3}{3} \times 25 + \frac{1}{3} \times 28 \approx 69.33$; value of $X^* = \frac{5}{3} \times 20 + \frac{3}{3} \times 30 + \frac{1}{3} \times 40 \approx 76.66$; value of $Y^* = \frac{5}{3} \times 21 + \frac{3}{3} \times 28 + \frac{1}{3} \times 40 \approx 76.33$.

[23] Compare Ebert (1988) and Asheim & Zuber (2014).

It is easy to see that generalized Ginis satisfy independence of the worst-off. The wellbeing of the worst-off, provided they remain worst-off and unconcerned, is going to cancel out on both sides when the weighted sums of the two outcomes are compared.

Analogously, weighted egalitarians will accept

> **Independence of the (Unconcerned) Best-Off**. If the best-off person is unconcerned between X and Y, and between X^* and Y^*, and the remaining people are equally well-off in X as in X^*, and in Y as in Y^*, then X is at least as good as Y iff X^* is at least as good as Y^*.

But why should weighted egalitarians accept independence of the worst-off and the best-off? Recall that weighted egalitarians do not care about inequality as such. Instead, they give importance to people's wellbeing based on how they relate to each other. In particular, whether a gain to the worse-off is worth a loss to the better-off depends on their "distance" from each other on the society's wellbeing ladder. It matters, for example, whether Cat is one or two rungs on the society's wellbeing ladder above Bob. Intuitively, that should be the only thing that matters for weighted egalitarians, besides wellbeing.

Now note that this motivation generalizes beyond the two independence principles just considered. Independence of the worst-off and the best-off means that the *wellbeing* of the worst-off and the best-off can be ignored if they are not affected. Can their *existence* also be ignored? Intuitively, for weighted egalitarians, removing the bottom or the top of the society's ladder of wellbeing should not affect how people in the middle compare, provided that their relative ranks are also unaffected. It is therefore natural for weighted egalitarians to accept

> **Existence Independence of the (Unconcerned) Worst-Off**. If the worst-off person in X and Y is unconcerned between X and Y and does not

exist in both X^* and Y^*, and the remaining people are equally well-off in X as in X^*, and in Y as in Y^*, then X is at least as good as Y iff X^* is at least as good as Y^*.

And

> **Existence Independence of the (Unconcerned) Best-Off.** If the best-off person in X and Y is unconcerned between X and Y and does not exist in both X^* and Y^*, and the remaining people are equally well-off in X as in X^*, and in Y as in Y^*, then X is at least as good as Y iff X^* is at least as good as Y^*.

As an extra advantage, accepting these two principles allows weighted egalitarians to say that inequalities between the present and future people, on the one hand, and the people in the far past, on the other hand, do not matter as long as the former are better-off than the latter. So, if we are certain everyone in the past was worse-off than everyone in the present or the future, the wellbeing level and the number of the far past people cannot influence our current redistributive policies. Research into the Inca civilization will not be typically relevant to current trade-offs between the worse-off and the better-off.[24] But what if we are not certain that the present and future people are better-off than those in the far past? Then it seems fine to let their position relative to the ancient Incas affect current trade-offs between them. Intuitively, if our contemporary is worse-off than an Inca peasant, say, that gives benefitting them a special urgency. Hence, weighted egalitarians should find these existence independence principles appealing.

---

[24] Fleurbaey (2010: 667-668) makes a similar point in favour of the generalized Gini family. It can allow egalitarians to escape another of Parfit's (2000 [1995], 1997) objections.

Still, the two principles significantly narrow down the sorts of weights that can be used in the generalized Gini formula. This is proven in full generality in the appendix; I will now consider some suggestive examples of generalized Ginis ruled out by the two existence independence principles.

Let us start with existence independence of the best-off, and consider how the following four outcomes are ranked by the standard Gini.[25]

	Ann	Bob	Cat	Gini
X	40	20	30	≈ 76.67
Y	40	21	28	≈ 76.33
X^*	Ω	20	30	45
Y^*	Ω	21	28	45.5

Table 3-5

First, to obtain Y from X, we make the same transfer from Cat to Bob as in Table 3-4: we take two units of wellbeing from Cat and give one unit of wellbeing to Bob, with Ann being better-off than both and unaffected by the transfer.

Now consider removing Ann from the compared outcomes, thus obtaining X^* and Y^*. Existence independence of the best-off implies that X is ranked against Y in the same way that X^* is ranked against Y^*.

[25] Calculations. Value of $X = \frac{5}{3} \times 20 + \frac{3}{3} \times 30 + \frac{1}{3} \times 40 \approx 76.67$; value of $Y = \frac{5}{3} \times 21 + \frac{3}{3} \times 28 + \frac{1}{3} \times 40 \approx 76.33$; value of $X^* = \frac{3}{2} \times 20 + \frac{1}{2} \times 30 = 45$; value of $Y^* = \frac{3}{2} \times 21 + \frac{1}{2} \times 28 = 45.5$.

But according to the standard Gini, X is better than Y yet X^* is worse than Y^*. That is, it is better to make the transfer when Ann is unaffected with forty units but worse to make the same transfer when Ann does not exist at all.[26]

This is because the standard Gini's weights depend on population size in a peculiar manner. For example, the lowest two ranks in a population of three receive weights $\frac{5}{3}$ and $\frac{3}{3}$, respectively, while the lowest two ranks in a population of two receive weights $\frac{3}{2}$ and $\frac{1}{2}$, respectively.

This means that the ratio of Cat's weight to Bob's weight in our example changes depending on the size of the population to which they belong, even though their relations to each other do not change: Cat is still one rung above Bob on the society's ladder of wellbeing. And, so, the transfer from Cat to Bob looks more favourable when they are bottom and second-to-bottom in a two-person outcome than if they are bottom and second-to-bottom in a three-person outcome.

Let us now consider an example of a generalized Gini which satisfies existence independence of the best-off but violates existence independence of the worst-off. The following formula (stated only for the case of three people) represents what we might call the "harmonic Gini":[27]

$$\text{Value} = w_1 + \frac{1}{2}w_2 + \frac{1}{3}w_3 \tag{7}$$

Consider how the harmonic Gini ranks the following four outcomes.

---

[26] A similar case works for Adler's example of a weighted-egalitarian theory, where weights are given simply by consecutive integers, with the highest integer weighing the worst-off person's wellbeing. See Adler (2012: 353).

[27] In general, the harmonic Gini draws its weights from the harmonic sequence: $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, ...$

	Ann	Bob	Cat	Harmonic Gini
X	10	20	30	30
Y	10	23	25	≈ 29.83
X^*	Ω	20	30	35
Y^*	Ω	23	25	35.5

Table 3-6

First, to obtain Y from X, we again make a transfer from Cat to Bob, but this time the transfer is more leaky than the one considered in Table 3-4 and Table 3-5. Here the transfer consists in taking five units of wellbeing from Cat and giving three units of wellbeing to Bob. Ann is worse-off than both and unaffected by the transfer.

Now consider removing Ann from the compared outcomes, thus obtaining X^* and Y^*. Existence independence of the worst-off implies that X is ranked against Y in the same way that X^* is ranked against Y^*.

But according to the harmonic Gini, X is better than Y while X^* is worse than Y^*. That is, it is better to make the transfer when Ann exists with ten units of wellbeing but worse to make the same transfer when Ann does not exist at all.[28]

While the harmonic Gini uses the same sequence of weights at all population sizes, these weights are not related to each other in the right way, relative to each population size. In particular, the ratio of Cat's weight to Bob's weight changes depending on the number of people worse-off than them, even though their relations to each other do not change: Cat is still one rung above Bob on the society's ladder of wellbeing. And, so, the transfer from Cat to Bob looks more

[28] Calculations. Value of $X = 1 \times 10 + \frac{1}{2} \times 20 + \frac{1}{3} \times 30 = 30$; value of $Y = 1 \times 10 + \frac{1}{2} \times 23 + \frac{1}{3} \times 25 \approx 29.83$; value of $X^* = 1 \times 20 + \frac{1}{2} \times 30 = 35$; value of $Y^* = 1 \times 23 + \frac{1}{2} \times 25 = 35.5$.

favourable when they are bottom and second-to-bottom in a two-person outcome than if they are second-to-bottom and third-to-bottom in a three-person outcome.

In this way, two natural members of the Gini family, the standard Gini and the harmonic Gini, are ruled out by the two existence independence principles that, I argued, weighted egalitarians have reason to accept.

These two examples generalize. As proved and explained in the appendix, weights used in the generalized Gini formula have to form a geometric sequence, with the same common ratio across all population sizes. This common ratio is effectively a discount factor and the weights act as discounts on higher ranks. The two existence independence principles leave open the choice of the initial weight at each population size. But it is natural to set that to unity. We then obtain the geometric Gini as the right generalized Gini to use in same-number cases.

4 The geometric Gini in different-number cases

The previous section argued that out of all the possible generalized Ginis, weighted egalitarians should use the geometric Gini in same-number cases. That completes the first horn of the dilemma from the introduction: weighted egalitarians can only reject the geometric Gini by rejecting existence independence principles they should find appealing. This section starts building towards the second horn of the dilemma: showing that the geometric Gini is implausible and unmotivated in different-number cases.[29]

[29] One could object that weighted egalitarians who accept the geometric Gini in same-number cases could extend it in some other way to different-number cases, much like proponents of the prioritarian formula in same-number cases can combine it with a "totalist" or an "averagist" aggregation method when it comes to different-number cases. Compare Holtug (2007) and

But let us start with some good news about the geometric Gini in different-number cases. First off, the geometric Gini avoids

> **The Repugnant Conclusion**. Any perfectly equal outcome is worse than some perfectly equal outcome where average wellbeing is arbitrarily close to zero.[30]

This is because, as Asheim & Zuber (2014) put it, the geometric Gini takes each wellbeing level to be of bounded importance in the sense that it evaluates an outcome where everyone has wellbeing level w at:

$$(1 + \beta + \beta^2 + \cdots) \times w \tag{8}$$

Since the discount factor β is between zero and one, this formula multiplies the wellbeing level w by a convergent geometric series which cannot exceed $1/(1-\beta)$. For example, if we set β to 0.8, no outcome consisting solely of people at wellbeing level w can exceed the value of $1/(1-0.8)w = 5w$. Or, to put it differently, a five-fold decrease in average wellbeing cannot be compensated by any increase in the number of people.

Second, the geometric Gini also satisfies

Arrhenius (2009). This is a possible way out, as far as this paper is concerned. Note, however, that it is the method of aggregation that is supposedly distinctive of weighted egalitarianism. Besides, going for "averagism", say, would conflict with the limited existence independence principles that weighted egalitarians should accept.

[30] Compare Parfit (1984: 388)

> **Priority for Lives Worth Living**. A perfectly equal outcome where everyone has positive wellbeing is better than a perfectly equal outcome where everyone has negative wellbeing.[31]

This is because a weighted sum all of whose terms are positive is bound to be greater than a weighted sum all of whose terms are negative, given that the weights are all positive, too.

But the geometric Gini fails to satisfy

> **The Mere Addition Principle**. Adding people with positive wellbeing without affecting the original people's wellbeing is at least as good as not adding them.[32]

This is because, according to the geometric Gini, creating a person has two effects on value. The added person is to be found at some rank in the expanded outcome. Adding them therefore raises the ranks of everyone who used to be at that rank or above. This effect is negative insofar as the people whose ranks are pushed upwards had positive wellbeing, and positive otherwise. The second effect has to do with the added person's own wellbeing: this effect is positive if the added person has a life worth living and negative otherwise.

The mere addition principle can fail precisely because the first effect might outweigh the second, as in the following example.

---

[31] See Blackorby et al. (2005: 135, 165). As they note, satisfying this priority axiom implies avoidance of Arrhenius's (2000) *strong* sadistic conclusion.

[32] Compare Parfit (1984: chapter 19) and Ng's (1989) reconstruction of Parfit's argument. The formulation in the main text replaces their "not worse" with "at least as good", which is harmless here.

	Ann	Bob	Geometric Gini
X	100	10	90
Y	100	Ω	100

Table 3-7

Setting the discount factor β to 0.8 again, the geometric Gini recommends Y over X. This is an example of what we can call "levelling-out": we can increase value by making sure the worst-off never come into existence. Since levelling-out removes bad inequality, there might be egalitarian reason to think levelling-out can be good all things considered.[33]

The three implications noted so far are often considered desirable in a theory of population ethics.[34] But the geometric Gini has some unappealing implications, too. For example,

> **The Sadistic Conclusion**. When adding people without affecting the original people's wellbeing, it can be better to add some people with negative wellbeing rather than some (possibly different) people with positive wellbeing.[35]

To see this, consider the following three outcomes.

[33] See Parfit (1984: section 144). Parfit claimed that levelling-out can never be deemed better on egalitarian grounds, not even better in any respect. As Temkin (2012: 383) reports, Parfit came to think this was his "worst mistake in philosophy". Persson (2001: 35) seems to claim (mistakenly) that a view like the generalized Gini can never recommend levelling-out.

[34] They are cited in support of a version of the geometric Gini by Asheim and Zuber (2014).

[35] See Arrhenius (2000).

	Ann	Bob	Dan	Cat1	…	Cat6	Geometric Gini
X	1	10	Ω	Ω	…	Ω	9
Y	1	10	Ω	1	…	1	≈ 6.05
Z	1	10	–1	Ω	…	Ω	6.2

Table 3-8

We can think of X as the status quo. Y results from adding six people (the Cats) with one unit of wellbeing each, while Z results from instead adding a single person (Dan) with one negative unit of wellbeing. Setting the discount factor β to 0.8 again, the geometric Gini implies that Z is better than Y.[36] So, it is better to add one miserable Dan than six content Cats.

The sadistic conclusion is a consequence of the geometric Gini for much the same reasons as the mere addition principle is not. Adding the six Cats brings in extra positive wellbeing but also decreases the weight of Bob's large contribution. Adding Dan brings in extra negative wellbeing but does not do as much to decrease the weight of Bob's contribution. In this way, the effect on other people's ranks can be what counts against creating a happy life.[37]

---

[36] Calculations. Value of X=$1 + 0.8 \times 10 = 9$; value of Y= $1 + 0.8 \times 1 + 0.8^2 \times 1 + 0.8^3 \times 1 + 0.8^4 \times 1 + 0.8^5 \times 1 + 0.8^6 \times 1 + 0.8^7 \times 10 \approx 6.05$; value of Z= $-1 + 0.8 \times 1 + 0.64 \times 10 = 6.2$.

[37] Asheim and Zuber (2014) use a "prioritarian" version of the geometric Gini where wellbeing levels are transformed by a strictly increasing, strictly concave transformation function before they are weighted by weights which depend on people's ranks. They show that, if this transformation function is bounded above, their version of the geometric Gini can satisfy the *non-sadism condition* which says that there is some negative wellbeing level such that adding a group of people at that level is at least as bad as adding any other group of people at a positive wellbeing level. But avoidance of the sadistic conclusion seems more natural as an axiom than the weak non-sadism condition. Moreover, the use of prioritarian transformations seems foreign to the core weighted-egalitarian idea.

While it is natural to reject the mere addition principle on egalitarian grounds, it is hard to find anyone who suggested embracing the sadistic conclusion on such grounds.[38] And it is even harder to make such a suggestion work in the present example. This is because, if anything, Y appears to be less unequal (and more equal) than Z. We can reach this conclusion in multiple ways. First, the relative deviation from the average is less in Y than in Z. Second, the Gini index is less for Y than for Z. Lastly, in Y the number of relations of equality is 21 and the number of relations of inequality is 7, while in Z it is 0 and 3, respectively.

Hence, we see that while the geometric Gini has some advantages in different-number cases, it also has some serious costs. The next section will focus more specifically on the dubious egalitarian credentials of the geometric Gini in different-number cases. This will complete the second horn of the dilemma against weighted egalitarianism, showing that, besides being implausible, the geometric Gini is ill-motivated as an egalitarian theory.

5 Is the geometric Gini egalitarian?

In this section, I will first show that the geometric Gini dampens people's wellbeing when this has no effect on inequality. Second, I will show that, according to the geometric Gini, it is better if a fixed positive total of wellbeing is divided amongst fewer people and that a fixed negative total of wellbeing is divided amongst more, even though in both cases there is no effect on inequality. I will also show that the geometric Gini cannot be justified by the veil of ignorance.

[38] But see Arrhenius (2013) and Segall (2019) where this idea is given serious consideration.

I then consider two responses to our argument: one trying to offer an alternative egalitarian justification of these features of the geometric Gini, the other modifying the formula for the geometric Gini in a way which avoids them.

5.1 Diminishing marginal value of extra people

First, the geometric Gini implies that extra people at a given wellbeing level have "diminishing marginal value". This is both odd and hard to square with egalitarian thinking.

To see this, consider the following two outcomes.

	Ann	Bob	Cat	Geometric Gini
X	100	100	Ω	180
Y	100	100	100	244

Table 3-9

The only difference between X and Y is that there is an extra person with 100 units of wellbeing in Y. Both X and Y are perfectly equal: everyone who exists has the same wellbeing level. But setting the discount factor β to 0.8, we see that the geometric Gini implies that Cat's contribution to value is merely 64 units rather than 100 units. In this way extra people count for less because we already have a number of people who are just as well-off.

This is hard to explain on egalitarian grounds, if we accept the following two plausible egalitarian principles. The first is that the value contribution of a person should be equal to their wellbeing unless there is some effect on the inequality of the outcome. The second is that any two completely equal outcomes are equally unequal. Together they imply that Cat cannot contribute less than 100 units of wellbeing in the move from X to Y, contrary to what the geometric Gini implies.

It is easy to see that the communal-egalitarian formula according to which the value of an outcome is just total wellbeing minus the badness of inequality has to declare X and Y equally good, since no bad inequality is present in either X or Y. So, unlike weighted egalitarians, all communal egalitarians can accept Persson's principle that "if all individuals are justly equally well-off, the moral value of the outcome equals the value of the sum of well-being that the outcome contains…" (2008: 298). Communal egalitarians need not think that a person's wellbeing can contribute less to the value of an equal outcome merely because there are other people who are just as well-off.

The behaviour of the geometric Gini in our current example is odd for another reason: it resembles the behaviour of so-called *variable-value theories* which can be represented by means of the following formula:[39]

$$\text{Value} = f(n) \times \text{average wellbeing} \tag{9}$$

where n is the number of people and f is an increasing, bounded and, hence, concave function.

Indeed, the geometric Gini takes the form of a variable-value theory when applied to perfectly equal outcomes: the value of n people at some wellbeing level w is given by the sum of the first n terms of a geometric sequence times w. Because the discount factor in the geometric Gini is between one and zero, the sum of the first n terms of a geometric sequence is an increasing, bounded and concave function, just like function f in the formula above.

The degree to which an extra person at a given wellbeing level is dampened depends on the discount factor β which fixes the geometric Gini's weights and, hence, determines its attitude to inequality in wellbeing in same-number cases. In

[39] See Hurka (1983), Hudson (1987), Ng (1989), Sider (1991).

this way, the degree of priority given to the relatively worse-off determines the speed at which the value of extra people diminishes. But with completely equal outcomes there is no inequality to be taken account of. The geometric Gini therefore makes attitudes to inequality relevant in the wrong place.

It is therefore hard to see the geometric Gini as simply aggregating people's wellbeing in a way which gives priority to the relatively worse-off, without regard for impersonal bads such as the badness of inequality. Instead, by implying diminishing marginal value of extra people at a fixed wellbeing level, the geometric Gini seems additionally to give weight to the value of variety, in much the same way as variable-value theories do. For example, Hurka (1983: 497) thinks we give a greater non-instrumental importance to conservation efforts aimed at whooping cranes than at common pigeons. The marginal whooping crane counts for more than the marginal pigeon, since the latter are much more numerous. In a similar way, the more people at a given wellbeing level there are, the less an extra one counts in terms of the geometric Gini. This is hard to square with the motivation behind weighted egalitarianism.

5.2 Preferences regarding fragmentation

A second feature of the geometric Gini which is both odd and hard to square with egalitarian thinking concerns what I will call "fragmentation".[40] One's attitude to fragmentation is revealed by one's answer to the following question: "If we have

[40] This label is due to Blackorby et al. (2005: 151-2) who also provide useful discussion. Broome's (2004: 108-109) discussion of the value of longevity is related, but fragmentation is not really about longevity, since our examples do not even mention the length of people's lives.

to divide a fixed total of wellbeing amongst some number of people, which number is it best to choose?"

If the total is positive, the geometric Gini prefers to choose as small a number of people as possible. It is therefore against fragmentation. As an example, consider the following two outcomes.

	Ann	Bob	Cat	Geometric Gini
X	100	Ω	Ω	100
Y	Ω	50	50	90

Table 3-10

Set the discount factor β to 0.8 as before. Then the geometric Gini recommends X over Y. It is better that one person has 100 units of wellbeing than if two people have 50 units each.

The geometric Gini's preference changes direction, however, once the total of wellbeing to be distributed is negative. For example, in the following table, the geometric Gini with the same discount factor as before recommends W over Z. It is better than two people have –50 units of wellbeing each rather than that one person has –100 units.

	Ann	Bob	Cat	Geometric Gini
Z	–100	Ω	Ω	–100
W	Ω	–50	–50	–90

Table 3-11

We could say that, according to the geometric Gini, happiness is to be spread as thickly as possible and misery is to be spread as thinly as possible.

This feature is problematic for the same sorts of reasons as the "diminishing marginal value of extra people" implied by the geometric Gini. First, the

geometric Gini's preferences regarding fragmentation are hard to explain on egalitarian grounds, since all four outcomes at issue are perfectly equal. This is another case where some weighted egalitarians will disagree with all communal egalitarians. Second, it is odd that one's desire to give priority to the relatively worse-off should imply a preference for or against fragmentation. The geometric Gini makes inequality aversion relevant in the wrong place.[41]

5.3 Replication invariance

To explain the third noteworthy feature of the geometric Gini, I need more terminology. Let the *n-replication* of outcome X be an outcome consisting of n copies of X, so that, if X has one person at level w, the n-replication of X has n people at w.

Then we can see that the geometric Gini violates the condition of

[41] The behaviour of the geometric Gini should be compared with that of prioritarianism and critical-level utilitarianism in different-number cases. See the discussion of fragmentation in Blackorby et al. (2005) cited above. Prioritarians (provided that their "transformation functions" are defined for negative amounts of wellbeing) will prefer fragmentation on the positive side and disprefer it on the negative side. Critical-level utilitarians will disprefer it on both sides. Prioritarianism is the mirror image of the geometric Gini in that respect. The geometric Gini implies the same kind of preferences regarding fragmentation as variable-value theories, including average utilitarianism. It is easy to see that according to average utilitarianism, it is also true that happiness is to be concentrated and misery is to be dispersed.

Replication Invariance. Outcome X is at least as good as outcome Y iff the n-replication of X is at least as good as the n-replication of Y, provided that X and Y are of the same size.[42]

This is a consistency condition about how comparisons at one population size relate to comparisons at a different size. It is violated by the geometric Gini in the following example.

	People		
	Ann	Bob	Geometric Gini
Outcome X	1	20	17
Outcome Y	2	18	16.4

	People								
	Ann1	Ann2	Ann3	Ann4	Bob1	Bob2	Bob3	Bob4	Geo. Gini
Outcome Z	1	1	1	1	20	20	20	20	≈ 27.13
Outcome W	2	2	2	2	18	18	18	18	27.67

Table 3-12

Compared with X, Y makes a leaky transfer from Bob to Ann: Bob loses two units of wellbeing and Ann gains one unit. Outcomes Z and W are 4-replications of outcomes X and Y, respectively. Replication invariance therefore requires that X is at least as good as Y iff Z is at least as good as W. But if we set the discount factor β to 0.8, we can see that, according to the geometric Gini, X is better than Y while Z is worse than W.[43]

[42] See Blackorby et al (2005: 123). This is related to what Ebert (1988) and Donaldson & Weymark (1980) call "Dalton's principle of population", originally suggested by Dalton (1920: 357).

[43] Calculations. Value of $X = 1 + 0.8 \times 20 = 17$; value of $Y = 2 + 0.8 \times 18 = 16.4$; value of $Z = \frac{1-0.8^4}{1-0.8} + 0.8^4\left(\frac{1-0.8^4}{1-0.8}\right) \times 20 \approx 27.13$; value of $W = \frac{1-0.8^4}{1-0.8} \times 2 + 0.8^4\left(\frac{1-0.8^4}{1-0.8}\right) \times 18 \approx 27.67$.

Why does this happen? As the outcomes X and Y are replicated, the collective weight received by the worse-off makes up an ever greater proportion of the total weight to be distributed across the whole population. It is about 56% when the population size is two and about 71% when the population size is eight.[44] As the population size increases to infinity, this proportion increases tends to 100%. In the limit, the worst-off are responsible for 100% of the value of the outcome.[45]

In our current example this translates into the fact that taking two units each from the best-off is not compensated by a one-unit each benefit to the worse-off when there is one of each in the best-off and worst-off bracket, but is compensated if there are four people in each bracket. This is odd.

It is also a problem if weighted egalitarianism is to be justified by the veil of ignorance, as in section 2, since any theory which violates replication invariance cannot be justified in that way.[46] To see this, consider the following four prospects.

---

[44] Calculations. First proportion $= \frac{1}{1+0.8} \approx 0.56$. Second proportion $= \frac{1-0.8^4}{1-0.8} \div \frac{1-0.8^8}{1-0.8} = \frac{1-0.8^4}{1-0.8^8} \approx 0.71$.

[45] This feature of the geometric Gini is noted by Fleurbaey, Tungodden & Vallentyne (2009: 276), Fleurbaey & Tungodden (2010: 405), and Asheim and Zuber (2014: 635-6).

[46] This point is due to Thomas (2016: 131) and Thomas et al. (2020: Proposition 6.1).

	States of nature	
Probabilities	1/2	1/2
Prospect X^*	1	20
Prospect Y^*	2	18

	States of nature							
Probabilities	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8
Prospect Z^*	1	1	1	1	20	20	20	20
Prospect W^*	2	2	2	2	18	18	18	18

Table 3-13

These prospects correspond to the four outcomes in Table 3-12. While prospect Z^* is defined over eight states of nature rather than just two, it offers the same probabilities of the same wellbeing levels as prospect X^*. Analogously for prospects W^* and Y^*.

Plausibly, all that matters in evaluating such prospects should be simply quantities of wellbeing and their probabilities. But then it follows that one should be equally well-off with X^* as with Z^*, and with Y^* as with W^*. Since the individual betterness relation on prospects is transitive, it follows that X^* is at least as good for one as Y^* iff Z^* is at least as good for one as W^*.

The veil of ignorance principle then implies that, in Table 3-12, X is at least as good as Y iff Z is at least as good as W, contradicting the implication of the geometric Gini. So, the geometric Gini is incompatible with the veil of ignorance principle, which we saw in section 2 to be the key premise in the veil of ignorance argument for weighted egalitarianism. Insofar as weighted egalitarians have to accept the geometric Gini, their view cannot be justified by appealing to the veil of ignorance.

5.4 Two responses

There are therefore important reasons to doubt that the geometric Gini is a well-motivated egalitarian view. I will now consider two responses.

The first response appeals to a distinction between positive and negative egalitarianism. According to *negative egalitarianism* (or anti-inegalitarianism), it is better to have less inequality, other things equal. According to *positive egalitarianism* (or affirmative egalitarianism), it is better to have more equality, other things equal.[47]

One might think that positive egalitarianism can be used to support some of the problematic features of the geometric Gini. As we saw with diminishing marginal value of extra people and with preferences regarding fragmentation, the geometric Gini implies preferences between situations where there is no effect on total wellbeing or inequality, which might seem odd given anti-inegalitarianism but sensible given positive egalitarianism, as in these cases there is an effect on *equality*.

However, the direction of preference predicted by positive egalitarianism does not match the direction predicted by the geometric Gini. Positive egalitarianism is naturally taken to imply that creating another person at some positive wellbeing level is good in two ways: it creates extra wellbeing and creates new relations of equality to all the people who already exist at that wellbeing level. So, the contribution of such a person should be magnified. But we saw above that according to the geometric Gini it is instead dampened. Similarly, the geometric

[47] Parfit (2000 [1995]) thought this distinction "pedantic" (page 86). But Persson (2001, 2008) and Arrhenius (2013) made a good case for its importance.

Gini implies that happiness is to be concentrated while misery is to be spread out. Positive egalitarianism, on the other hand, is naturally taken to imply that both happiness and misery are to be spread out: by dividing a given total of wellbeing amongst a greater number of people we increase the number of valuable relations of equality. We should doubt whether the geometric Gini can be seen as egalitarian, whether positive-egalitarian or negative-egalitarian.

The second response is that I misrepresented the formula that weighted egalitarians use for evaluating outcomes. The problems described above do not arise if all people at the same wellbeing level are given the same weights. One might think that weighted egalitarians should reformulate the geometric Gini formula accordingly.

To do this, divide people in an outcome into *wellbeing brackets* so that people in the same wellbeing bracket are equally well-off. Order the brackets from worst-off to best-off. Let $\mathfrak{w}_i$ be the shared wellbeing level in the i-th worst-off bracket, and let $\mathfrak{n}_i$ be the number of people in that bracket. Then we can modify the geometric Gini formula as follows:

$$\text{Value} = (\mathfrak{n}_1\mathfrak{w}_1) + \beta(\mathfrak{n}_2\mathfrak{w}_2) + \beta^2(\mathfrak{n}_2\mathfrak{w}_2) + \cdots + \beta^{n-1}(\mathfrak{n}_n\mathfrak{w}_n) \tag{10}$$

It can easily be seen that this formula avoids all the odd consequences of the geometric Gini that I described in this section: it does not dampen the contribution of extra people at a given wellbeing level, it has no preferences regarding fragmentation, and it satisfies replication invariance.

Nonetheless, weighted egalitarians cannot accept this modification, since it implies that levelling-down can be all things considered better. To see this, consider the following two outcomes.

	Ann	Bob	Modified geometric Gini
X	120	100	196
Y	100	100	200

Table 3-14

Setting the discount factor β to 0.8 as before, the modified geometric Gini implies that Y is better than X. So bringing down Ann's wellbeing to Bob's level is an overall improvement.

This is because raising Ann's wellbeing by 20 units has two effects on the value of the outcome. The first effect is positive: there are 20 units more of wellbeing in aggregate. The second effect is negative: the 100 units that Ann already had now make a smaller contribution to overall value since they are discounted by a higher weight, reflecting Ann's advancement into the ranks of the best-off.

Part of the motivation for weighted egalitarianism was to avoid the thought that levelling-down can be good in some way. The modified geometric Gini implies that levelling-down can be good all things considered. So, weighted egalitarians cannot accept the proposed modification.

6 Conclusion

Egalitarians have to choose between weighted and communal versions of their view. This paper argued that weighted egalitarians are in no position to reject the geometric Gini. This leads to trouble, however, since there are features of the geometric Gini which seem odd and hard to square with egalitarian thinking. This dilemma suggests egalitarians should be communal egalitarians, after all, which means they will have to forgo the purported benefits of weighted egalitarianism. As far as egalitarianism goes, we are back to square one.

7 Appendix

I will now show how to use existence independence principles to narrow down the generalized Ginis to just the geometric Gini. The results concerning principles of existence independence are related to those in Ebert (1988) and can also be found in Asheim & Zuber (2014). The arguments to follow are new and more accessible, however.

7.1 Generalized Ginis

We need to determine how the weights in the generalized Gini formula (2) relate to each other. To begin with, the sequences of weights for each population size will be unconstrained, except for being positive and decreasing. "Weight k out of n" will denote the weight given to the k-th worst-off person in a population of n people, where $k \leq n$.

Suppose throughout the discussion to follow that Ann is a fixed number of ranks above Bob: to begin with, Ann is rank j and Bob is rank i in a population of n people, with $j > i$.

We will consider potentially leaky transfers from Ann to Bob, so that Ann's loss may be greater than Bob's gain. We also that assume all transfers keep everyone's ranks unchanged. We will say that a transfer is *break-even* iff it is just as good to make it as not.

According to the generalized Gini formula (2), a transfer from Ann to Bob is break-even iff

$$\frac{\text{gain to Bob}}{\text{loss to Ann}} = \frac{\text{weight } j \text{ out of } n}{\text{weight } i \text{ out of } n} \tag{11}$$

Why? The transfer is break-even iff the weighted sum of wellbeing of all ranks before the transfer is equal to the weighted sum of wellbeing of all ranks after the transfer. All ranks except for i and j are unaffected, so their weighted wellbeing cancels out, leaving us with:

$$(\text{weight } i \text{ of } n) \times (\text{gain to Bob}) = (\text{weight } j \text{ of } n) \times (\text{loss to Ann})$$

which is equivalent to (11).

Thus, according to the geometric Gini, the ratio of people's weights is the only factor which determines whether a transfer from the better-off to the worse-off is acceptable, provided that the transfer is rank-preserving.

7.2 Existence independence of the best-off

Now consider adding an unconcerned best-off person to the outcome, while leaving everything else the same. In that situation a transfer from Ann to Bob is break-even iff:

$$\frac{\text{gain to Bob}}{\text{loss to Ann}} = \frac{\text{weight } j \text{ out of } (n+1)}{\text{weight } i \text{ out of } (n+1)} \tag{12}$$

Existence independence of the best-off implies that whether a transfer is break-even cannot depend on whether this unconcerned best-off person exists. This means that (12) is true iff (11) is, for a fixed choice of Bob's gain and Ann's loss. And that means that the following identity is true:

$$\frac{\text{weight } j \text{ out of } n}{\text{weight } i \text{ out of } n} = \frac{\text{weight } j \text{ out of } (n+1)}{\text{weight } i \text{ out of } (n+1)} \tag{13}$$

So, for example, the ratio of weight 2 to weight 1 in a population of two people is the same as the ratio of weight 2 to weight 1 in a population of three people, and so on.

Since the weights are strictly decreasing, we can write each weight as *some fraction* of the previous weight: this fraction need not be the same for all pairs of weights.

But if weight 2 is some fraction p_1 of weight 1 when the population size is n, then it follows from (13) that weight 2 must be the same fraction p_1 of weight 1 when the population size is $(n+1)$, and so on. It follows we can picture weights at each population size as follows.

	Weights at size n	Weights at size m
Rank 1	u	v
Rank 2	$p_1 u$	$p_1 v$
Rank 3	$p_2 p_1 u$	$p_2 p_1 v$
Rank 4	$p_3 p_2 p_1 u$	$p_3 p_2 p_1 v$
…	…	…

Table 3-15

If we set the initial weights at each population size as equal to some constant, say, one, then we can write the generalized Gini formula (2) using the same sequence of weights for each population size. We can call this class of generalized Ginis the "single-series Ginis".[48]

[48] The class appears in Donaldson & Weymark (1980).

7.3 Existence independence of the worst-off

Now consider adding an unconcerned worst-off person to an outcome with Ann and Bob. The transfer from Ann to Bob is break-even after this addition iff:

$$\frac{\text{gain to Bob}}{\text{loss to Ann}} = \frac{\text{weight } (j+1) \text{ out of } (n+1)}{\text{weight } (i+1) \text{ out of } (n+1)} \tag{14}$$

Existence independence implies that whether a transfer is break-even cannot depend on whether this unconcerned best-off person exists. This means that (14) is true iff (11) is true, for a fixed choice of Bob's gain and Ann's loss. And that means that the following identity is true:

$$\frac{\text{weight } j \text{ out of } n}{\text{weight } i \text{ out of } n} = \frac{\text{weight } (j+1) \text{ out of } (n+1)}{\text{weight } (i+1) \text{ out of } (n+1)} \tag{15}$$

But we already established (13) on the strength of existence independence of the best-off, and, as an instance of (13), we obtain:

$$\frac{\text{weight } (j+1) \text{ out of } (n+1)}{\text{weight } (i+1) \text{ out of } (n+1)} = \frac{\text{weight } (j+1) \text{ out of } n}{\text{weight } (i+1) \text{ out of } n} \tag{16}$$

Now (15) and (16) together give us the following identity:

$$\frac{\text{weight } j \text{ out of } n}{\text{weight } i \text{ out of } n} = \frac{\text{weight } (j+1) \text{ out of } n}{\text{weight } (i+1) \text{out of } n} \tag{17}$$

This establishes a connection between weights relative to the same population size. So, for example, the ratio of weight 2 to weight 1 is the same as the ratio of weight 3 to weight 2, and so on.

Let β be the ratio of weight 2 to weight 1. Since the weights are decreasing but positive, β must be between 0 and 1. This means that $\frac{\text{weight 2}}{\text{weight 1}} = \beta$, so weight 2 = $\beta \times$ weight 1. Since the ratio of weight 3 to weight 2 is also β, it follows that $\frac{\text{weight 3}}{\text{weight 2}} = \beta$, so weight 3 = $\beta \times$ weight 2. Given how we just defined weight 2, this

means that weight $3 = \beta \times (\beta \times \text{weight } 1)$. We can carry on in the same manner, so that, in general, the sequence of weights at each population size can be written as follows.

	Weights at size n	Weights at size m
Rank 1	u	v
Rank 2	βu	βv
Rank 3	$\beta \times (\beta u)$	$\beta \times (\beta v)$
Rank 4	$\beta \times (\beta \beta u)$	$\beta \times (\beta \beta v)$
...	...	...

Table 3-16

If we set the initial weights at each population size as equal to some constant, say, one, then we can write the generalized Gini formula not just using a single sequence of weights across all population sizes (as in the single-series Gini), but by using a single geometric sequence with the common ratio β. This is the geometric Gini.[49]

[49] These results correspond to Ebert (1988: proposition 12) and Asheim & Zuber (2014: lemma 2).

8 References

Adler, M. (2012). *Well-being and fair distribution.* Oxford: Oxford University Press.

Anand, S. (1983). *Inequality and poverty in Malaysia: Measurement and decomposition.* New York: Published for the World Bank by Oxford University Press.

Arrhenius, G. (2000). An impossibility theorem for welfarist axiologies. *Economics and Philosophy, 16*(2), 247-266. doi:10.1017/S0266267100000249

Arrhenius, G. (2009). Egalitarianism and population change. In A. Gosseries, & L. Meyer (Eds.), *Intergenerational justice* (pp. 323-347) Oxford University Press.

Arrhenius, G. (2013). Egalitarian concerns and population change. In N. Eyal, S. A. Hurst, D. Wikler & O. F. Norheim (Eds.), *Inequalities in health* (pp. 74-91). Oxford: Oxford University Press.

Asheim, G. B., & Zuber, S. (2014). Escaping the repugnant conclusion: Rank-discounted utilitarianism with variable population. *Theoretical Economics, 9*(3), 629-650. doi:10.3982/TE1338

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory, 2*(3), 244-263. doi:10.1016/0022-0531(70)90039-6

Blackorby, C., Bossert, W., & Donaldson, D. J. (2005). *Population issues in social choice theory, welfare economics, and ethics.* Cambridge: Cambridge University Press.

Blackorby, C., & Donaldson, D. (1978). Measures of relative equality and their meaning in terms of social welfare. *Journal of Economic Theory, 18*(1), 59-80. doi:10.1016/0022-0531(78)90042-X

Broome, J. (2002). *Respects and levelling down.* Unpublished manuscript.

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

Brown, C. (2003). Giving up levelling down. *Economics and Philosophy, 19*(1), 111-134. doi:10.1017/S0266267103001044

Buchak, L. (2013). *Risk and rationality.* Oxford: Oxford University Press.

Buchak, L. (2017). Taking risks behind the veil of ignorance. *Ethics, 127*(3), 610-644. doi:10.1086/690070

Dalton, H. (1920). The measurement of the inequality of incomes. *The Economic Journal: The Quarterly Journal of the Royal Economic Society, 30*, 348-361.

d'Aspremont, C., & Gevers, L. (2002). Social welfare functionals and interpersonal comparability. In K. Arrow, A. Sen & K. Suzumura (Eds.), *Handbook of social choice and welfare* (pp. 459-541) Elsevier B.V. doi:10.1016/S1574-0110(02)80014-5

Donaldson, D., & Weymark, J. A. (1980). A single-parameter generalization of the Gini indices of inequality. *Journal of Economic Theory, 22*(1), 67-86. doi:10.1016/0022-0531(80)90065-4

Ebert, U. (1988). Measurement of inequality: An attempt at unification and generalization. *Social Choice and Welfare, 5*(2), 147-169. doi:10.1007/BF00735758

Fleurbaey, M. (2010). Assessing risky social situations. *Journal of Political Economy, 118*(4), 649-680. doi:10.1086/656513

Fleurbaey, M. (2015). Equality versus priority: How relevant is the distinction?*31*(2), 203-217. doi:10.1017/S0266267115000085

Fleurbaey, M., & Tungodden, B. (2010). The tyranny of non-aggregation versus the tyranny of aggregation in social choices: A real dilemma. *Economic Theory, 44*(3), 399-414. doi:10.1007/s00199-009-0462-0

Fleurbaey, M., Tungodden, B., & Vallentyne, P. (2009). On the possibility of nonaggregative priority for the worst off. *Social Philosophy and Policy, 26*(1), 258-285. doi:10.1017/S0265052509090116

Harsanyi, J. (1977). Morality and the theory of rational behavior. *Social Research, 44*(4), 623-656.

Hirose, I. (2004). Equality, priority, and aggregation. PhD thesis. University of St Andrews. Archived at: http://hdl.handle.net/10023/2690

Hirose, I. (2009). Reconsidering the value of equality. *Australasian Journal of Philosophy, 87*(2), 301-312. doi:10.1080/00048400802636395

Hirose, I. (2015). *Egalitarianism*. Abingdon: Routledge.

Holtug, Nils. 2007. 'On Giving Priority to Possible Future People.' In *Hommage À Wlodek: Philosophical Papers Dedicated to Wlodek Rabinowicz*, edited by Dan Egonsson, Jonas Josefsson, Björn Petersson, Toni Rønnow-Rasmussen and Wlodek Rabinowicz. Lund, Sweden: Department of Philosophy, Lund University.

Holtug, N. (2010). *Persons, interests, and justice*. Oxford: Oxford University Press.

Hudson, J. (1987). The diminishing marginal value of happy people. *Philosophical Studies, 51*(1), 123-137. doi:10.1007/BF00353967

Hurka, T. (1983). Value and population size. *Ethics, 93*(3), 496-507. doi:10.1086/292462

Maskin, E. (1978). A theorem on utilitarianism. *The Review of Economic Studies, 45*(1), 93-96. doi:10.2307/2297086

McCarthy, D., Mikkola, K., & Thomas, T. (2020). Utilitarianism with and without expected utility. *Journal of Mathematical Economics, 87*, 77-113. doi:10.1016/j.jmateco.2020.01.001

Ng, Y. (1989). What should we do about future generations? *Economics and Philosophy, 5*(2), 235-253. doi:10.1017/S0266267100002406

Otsuka, M., & Voorhoeve, A. (2018). Equality versus priority. In Olsaretti, S. (Ed.), *Oxford handbook of distributive justice* (pp. 65-85) Oxford: Oxford University Press.

Parfit, D. (1984). *Reasons and persons*. Oxford: Oxford University Press.

Parfit, Derek. 1991. *Equality Or Priority*. The Lindley Lecture: University of Kansas, Department of Philosophy. Retrieved from http://hdl.handle.net/1808/12405

Parfit, D. (1997). Equality and priority. *Ratio, 10*(3), 202-221. doi:10.1111/1467-9329.00041

Persson, I. (2001). Equality, priority and person-affecting value. *Ethical Theory and Moral Practice, 4*(1), 23-39. doi:10.1023/A:1011486120534

Persson, I. (2008). Why levelling down could be worse for prioritarianism than for egalitarianism. *Ethical Theory and Moral Practice, 11*(3), 295-303. doi:10.1007/s10677-007-9102-6

Rescher, N. (1966). *Distributive justice: A constructive critique of the utilitarian theory of distribution*. Indianapolis: Bobbs-Merrill.

Segall, S. (2019). Why we should be negative about positive egalitarianism.*31*(4), 414-430. doi:10.1017/S0953820819000219

Sen, A., & Foster, J. E. (1997). *On economic inequality* (Expanded edition ed.). Oxford: Clarendon Press.

Sider, T. R. (1991). Might theory X be a theory of diminishing marginal value? *Analysis, 51*(4), 265-271. doi:10.1093/analys/51.4.265

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature, 38*(2), 332-382. doi:10.1257/jel.38.2.332

Sugden, R. (1986). New developments in the theory of choice under uncertainty. *Bulletin of Economic Research, 38*(1), 1-24.

Temkin, L. S. (1993). *Inequality*. New York; Oxford: Oxford University Press.

Thomas, T. (2016). Topics in population ethics. DPhil thesis. University of Oxford. Archived at: https://ora.ox.ac.uk/objects/uuid:fa2a09aa-e784-4126-bd4a-0487d3653add

Weirich, P. (1983). Utility tempered with equality. *Noûs, 17*(3), 423-439. doi:10.2307/2215258

Chapter 4

People in Suitcases

Abstract: Ex-ante deontology seeks to combine standard deontic constraints on harming some to help others with the idea that acting in everyone's interest is always right. I show that this approach faces serious problems in cases where agents need to make multiple decisions across time. I then argue that these problems force us to choose between orthodox deontology and broadly consequentialist views, and that we should opt for the latter. I suggest how my argument vindicates appeals to veils of ignorance in ethics.[1]

Word count: 8520

[1] I would like to thank Tomi Francis, Korbinian Rüger, Jessica Fischer, Elliott Thornley, Alice Van't Hoff, Aidan Penn, and Todd Karhu. Special thanks to Ralf Bader, Teru Thomas, and Michal Masny for helpful written comments.

1 Introduction

In some cases acting in everyone's best interest means violating prohibitions on harming some to help others. For example, consider

> *Opaque Footbridge.* Aye, Bea and Cee are trapped in three suitcases. Two of them are on the railway track, in the path of an out-of-control lethal trolley. The other is on a footbridge above the track. No one knows who is where: the suitcases have just been shuffled. You can push a lever which will topple the footbridge suitcase onto the track, stopping the trolley from hitting the remaining two suitcases.[1]

We can represent your decision problem by means of the following decision tree.[2]

[1] This is a scaled-down version of Hare's (2016) case. See also Hare (2013: 89-96). Similar cases have been discussed by Thomson (1990: 176-202) and by many others typically under the heading of "survival lotteries". See Harris (1975) and Singer (1977), but also Kamm (1996: 143-171, 290-310). A real-life "survival lottery", proposed for Allied bombers during World War Two, is recounted by Glover (1977: 212-213).

[2] For a more detailed treatment of decision trees see Hammond (1988) and McClennen (1990: 99-111).

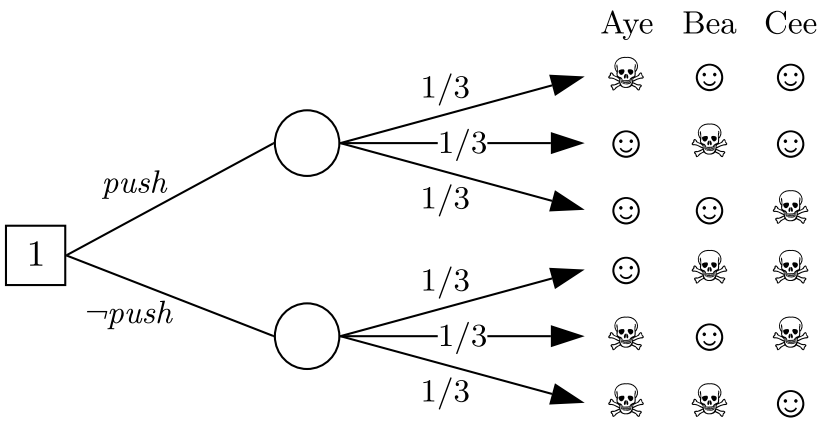


Figure 4-1. Opaque Footbridge

In this tree squares represent your choices, while circles represent moments when your uncertainty is resolved, with different possible resolutions having the probabilities shown. Aye, Bea and Cee can either die (denoted by a skull, ☠) or live (denoted by a happy face, ☺).

What should you do? There is a good case for pushing the lever: doing so is in everyone's interest. Pushing gives everyone a 1/3 chance of death and a 2/3 chance of life. Not pushing gives everyone a 2/3 chance of death and a 1/3 chance of life. If life is better than death, any rational self-interested person will rather face the former prospect.

But pushing means that someone will be killed so that two other people shall live. Many people have a strong intuition that harming some to help others is usually, if not always, forbidden. They believe in so-called *deontic constraints*.[3] They typically appeal to cases like

[3] See, for example, Nozick (1974: 28-33) and Kamm (1993, 1996, 2007). For objections and discussion, see Scheffler (1994 [1982]), Kagan (1989: 1-182), Otsuka (2011).

Transparent Footbridge. Aye, Bea and Cee are trapped in three suitcases. The situation is the same as in Opaque Footbridge except that everyone knows who is where.[4]

If it is Aye who is on the footbridge, we can represent that case as follows.

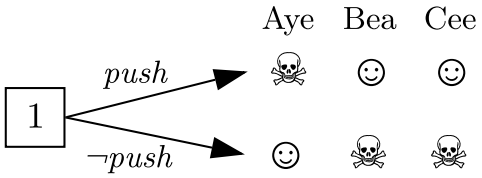


Figure 4-2. Transparent Footbridge

So, it seems that pushing is called for if the case is opaque but forbidden if the case is transparent. How should we resolve that tension? I see three main natural theoretical options, as in the following table.[5]

[4] This is a version Thomson's (1985) case.

[5] Compare Adler and Sanchirico's (2006: 289) remark: "Choice under uncertainty is a *general* issue for normative theories, not unique to welfarism, and close analogues to the ex ante/ex post problem arise within various nonwelfarist views. For example, (…) a theory that recognizes certain moral rights might specify those in ex post terms (a right not to be killed) or in ex ante terms (a right not to be put at high risk of death)". See also Fried (2012: 525-529).

	Opaque Footbridge		Transparent Footbridge	
	push	$\neg$*push*	*push*	$\neg$*push*
Permissivism	✔	×	✔	×
Ex-ante deontology	✔	×	×	✔
Orthodox deontology	×	✔	×	✔

Figure 4-3. Possible views

Let's start with the two extremes. The first is what I will call *permissivism*, which simply does away with deontic constraints. It implies that you should push in both cases. Permissivists are often consequentialists, but there are also forms of deontology which reject deontic constraints while holding on to *moral options*: permissions to do less than best.[6] The other extreme I will call *orthodox deontology*. It simply says that actions sure to harm some to help others are prohibited, barring very special circumstances. It implies that you should not push in either case.[7]

But there is also a range of views in the middle.[8] I think the most attractive one is what I will call

> **Ex-Ante Deontology**. Choose an action that gives everyone best prospects, even if, as a result, some people will be harmed to help others. Don't harm people in that way otherwise.[9]

---

[6] See Scheffler (1994 [1982]) and Kagan (1989: 183-203).

[7] Kamm (1996: 290-310) and Alexander (2014) seem to have that sort of view.

[8] Compare Thomson (1990: 176-202), Hare (2013, 2016), Frick (2015) and Setiya (2020). I will discuss these authors in more detail below.

[9] A more careful formulation would add a number of *ceteris paribus* clauses and exceptions about self-defence, war, punishment, medical consent, and so on. But it will do for my purposes. We can

It implies that you should push if people's whereabouts are unknown, but you should stay put if they are.

Ex-ante deontology can also be given a rationale that should be appealing to deontologists. To see this, recall that Nozick, a prominent champion of deontic constraints, claims that they "reflect the fact of our separate existences":

> "There are only individual people, different individual people, with their own individual lives. Using one of these people for the benefit of others, uses him and benefits others. Nothing more" (1974: 33).

The individual used "does not get some overbalancing good from his sacrifice, and no one is entitled to force this upon him". So, it looks like, for Nozick, deontic constraints are meant to protect individuals from an uncompensated sacrifice of their interests.

But, as we saw, in some cases, an action that uses some to benefit others is in everyone's interest, at least relative to what is known at the time of action. Nozick's comments suggest that in these cases deontic constraints are lifted. In other cases, individual interests conflict and, so, deontic constraints kick in.[10]

Nonetheless, in this paper I will argue against ex-ante deontology and other compromise views like it. In sections 2 to 4 I will show that ex-ante deontology faces serious problems in cases where the agent is called to make multiple choices over time. In section 5 I generalize the problem by deriving a veil-of-ignorance principle which rules out deontic constraints. Still, ruling out ex-ante deontology and other compromise views leaves us with a choice between permissivism and

---

see ex-ante deontology as trying to reconcile deontology with a deontic version of the principle of *ex-ante Pareto*, on which see, for example, Adler (2012: 477-551).

[10] Alternative explanations for deontic constraints have of course been offered. See Kamm (1996).

orthodox deontology. In section 6 I take up the cause of the former by formulating an argument for pushing in Opaque Footbridge. In section 7 I also point to new problems for deontologists who would like to act in everyone's interests at least in the limited range of cases where, unlike in Opaque Footbridge, the true state of nature is known in advance. I conclude that deontology is unable to show minimal respect to flesh-and-blood persons as opposed to persons in the abstract. That is a serious problem.

2 Problems for Ex-Ante Deontology

Ex-ante deontology runs into problems in cases which are like Opaque Footbridge except that, after deciding to pushing the lever you find out who is where and get a chance to change your mind before it's too late. As an example, consider

> *Case One.* Aye, Bea and Cee are trapped in three suitcases. The situation is the same as in Opaque Footbridge, except for the following details. It is now 12:00. You still don't know who is where but, luckily, everyone will peek out of their suitcases by 13:00. The trolley is slow and will only roll under the footbridge at 13:15. The lever is also rusty. You can use it to topple the footbridge suitcase but only if you start pushing right away and keep pushing until the trolley arrives. If you stop pushing at 13:00, then whoever is on the footbridge will see you and will become mildly traumatized, realizing they could easily have died a terrible death as a train stop. Alternatively, you can walk away right now and let the trolley run its course.

We can represent your decision problem as follows, with a sad face (☹) denoting mild psychological trauma.

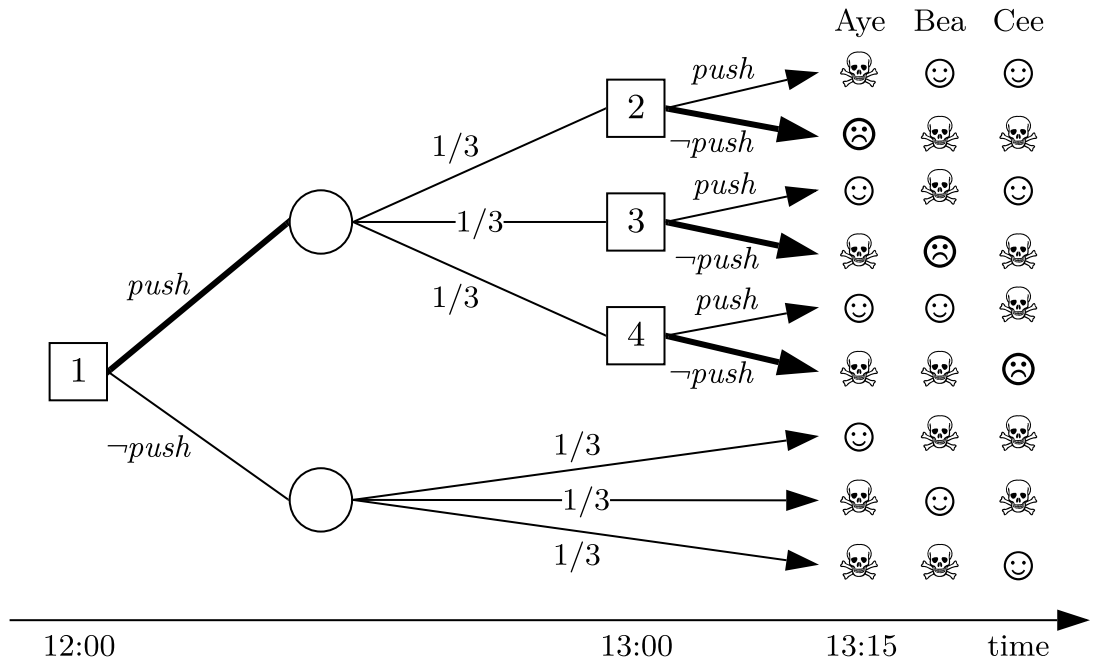


Figure 4-4. Case One

What should you do? The bold lines above show what I take to be ex-ante deontology's answer. Pushing is in everyone's interest at 12:00. So, ex-ante deontology implies that you should start to push at 12:00. In terms of the decision tree, this means going up at choice node 1.[11]

But at 13:00 pushing is no longer in everyone's interest. If it is Aye who is on the footbridge, for example, then pushing is not in Aye's interest, even though it benefits Bea and Cee. So, deontic constraints kick in and ex-ante deontology implies that you should not push the lever at 13:00. In terms of the decision tree, this means going down at choice nodes 2, 3, 4.

So, overall, ex-ante deontology tells you to push at 12:00 but then to bail out at 13:00. The result is two dead bodies on the track and one mildly traumatized

[11] If you suspect that ex-ante deontology can avoid this implication if combined with a suitable method of choice for sequential decision problems, hold fire until the next two sections.

person on the footbridge. If you simply walked away at 12:00, at least the footbridge person would have been happier. Why bother pushing the lever at all?

There are at least two problems for ex-ante deontology coming from that story. The first is that ex-ante deontology can force moral agents to predictably and pointlessly go back on their past decisions. That seems irrational. We can put this point by saying that ex-ante deontology violates

> **Dynamic Consistency**. Continuations of permissible actions are always permissible. Continuations of impermissible actions are always impermissible, provided no mistake was made in the meantime.[12]

The second problem is that this mandated fickleness can be costly. As we saw, ex-ante deontology can lead you to make sure that someone is gratuitously traumatized. That seems immoral.

Put more carefully, the outcome of deciding to push at 12:00 but then backtracking at 13:00 is certain to be exactly the same, in all relevant respects, as the outcome of not pushing at 12:00, *except* that some people are worse-off than they would have been. In both cases the trolley is allowed to run its course and no one is used as a train stop. If given a direct choice between the two sorts of outcomes, any sensible deontologist would choose the latter. Yet following ex-ante deontology means ending up with the former! This looks like the theory defeating itself.

We can put the point a bit more abstractly by saying that ex-ante deontology violates

---

[12] This is my formulation of McClennen's (1990: 120) condition of the same name. Why the "no mistake" proviso? Some alternatives may still be permissible if they become available, even if it is impermissible to make them available.

> **Sequential Dominance**. One cannot permissibly carry out a course of action that is certain to have the same outcome, in all relevant respects, as some other available course of action, except for someone being worse-off.[13]

The conflict with sequential dominance is especially awkward for ex-ante deontology. After all, the theory tries to preserve the idea that one should act in everyone's interest if possible. But here it ends up traumatizing people for nothing. Ex-ante deontology begins to look like an unstable hybrid between the purer extremes of permissivism and orthodox deontology.

I think Case One poses similar problems for other compromise that that try to combine standard deontological verdicts with acting in everyone's interest. I know two such views, one due to Hare (2013, 2016), the other due to Setiya (2020). Here is what they imply in Opaque and Transparent Footbridge.

	Opaque Footbridge		Transparent Footbridge	
	push	$\neg$*push*	*push*	$\neg$*push*
Hare's view	✔	×	✔	✔
Setiya's view	✔	✔	×	✔

Figure 4-5. Other compromise views

For both of them the contrast between opaque and transparent versions of Footbridge is less stark than for ex-ante deontology.

[13] Similar, albeit distinct, principle appears in Gustafsson (2015: 1594-1595). Another upshot of Case One is that ex-ante deontology makes you give everyone worse prospects than you could have. Compare Gustafsson (2018: 599).

Let's start with Hare's view. He thinks that pushing is the only right action in Opaque Footbridge, while both pushing and not pushing are permissible in Transparent Footbridge. Hare's explanation is that knowing who is where induces incommensurability between your reasons to push and not to push: your reasons to push are neither stronger than, nor weaker than, nor just as strong as, your reasons not to push. The needed extra knowledge is glossed as biographical knowledge: incommensurability sets in once we know more about the richly textured lives of the people affected. Hare adds that "faced with such incommensurability, you need a tiebreaker, and a policy of nonintervention in cases like this is as good a tiebreaker as any" (2016: 467).

I don't think Hare's view is very appealing. First, it makes deontic constraints into mere tie-breakers: something much less robust than deontologists typically have in mind. Second, it is unclear why knowing who is where induces incommensurability, especially given Hare's gloss on the sort of knowledge at stake. As Setiya puts it, it would not matter to permissibility "if the people involved were perfect duplicates of one another (…) who lead identical solitary lives. Nor would it matter if they were people you just met and about whom you know nothing at all" (2020: 70).[14]

But, more importantly, Hare's view does not avoid the problems posed by Case One. In that case, Hare's view *requires* pushing at 12:00 but then *allows* you to bail out at 13:00. So, all in all, it still *permits* you to gratuitously traumatize the person on the footbridge. It is therefore incompatible with both dynamic consistency and sequential dominance.

---

[14] Setiya seems unaware that Thomson (1990: 211-220, 185-186) apparently did think that it matters whether or not the people affected are clones. See Kamm (1993: 374-376) for discussion of Thomson's views.

Let's now move to Setiya. He argues that if pushing is mandatory in Opaque Footbridge, then it is also mandatory in Transparent Footbridge. Since he sticks to the standard deontological verdict in the latter case (don't push!), he thinks that pushing cannot be the only right course of action in the former. Instead, he suggests that both pushing and not pushing are permissible. The explanation here is that while pushing is in everyone's interests, not pushing is supported by people's "right to be free of unwanted intervention" (2020: 68). So, overall, in Opaque Footbridge there is incommensurability between your reasons to push and not to push. Setiya adds, however, that

> "Paternalistic intervention is easier to justify when it is the life of a relative or close friend in which you intervene. Relationships attenuate the boundaries between us, the claims of our autonomy" (2020: 68)

I think Setiya's view is no more appealing than Hare's. Just like the case of solitary clones is a problem for Hare, young children, coma patients, and friends and intimates are a problem for Setiya, as their autonomy is either nonexistent or attenuated by relationships to us. Just like Hare's view seems to require killing an identical solitary triplet to save his siblings, Setiya's view seems to require killing a friend to save two other friends or a child to save two other children, while upholding higher standards for autonomous strangers and adults. I don't think deontologists will find these implications appealing.

But, more importantly, Setiya's view also faces problems in Case One, where it *allows* pushing at 12:00 but *requires* bailing out at 13:00. So, much like Hare's

view, it is incompatible with both dynamic consistency and sequential dominance.[15]

[15] To be fair, neither Hare's nor Setiya's views *force* you to traumatize the footbridge person: they merely *allow* you to. I think that's bad enough. Compare the discussion of *non-forcing money pumps* in Gustafsson and Espinoza (2010) and Peterson (2015).

3 Sophisticated Ex-Ante Deontology

One obvious response to last section's argument is that, when deciding whether to push at 12:00, you should take into account whether you *would* also push at 13:00. This is the key idea behind the method of *sophisticated choice*: you have to hold fixed your future decisions when deciding what to do earlier.[16] We can revise ex-ante deontology correspondingly:

> **Sophisticated Ex-Ante Deontology**. Choose an action that gives everyone best prospects, *given what you are permissibly going to do in the future*, even if, as a result, some people will be harmed to help others. Don't harm people in that way otherwise.[17]

We can see how this works in the decision tree illustrating Case One, with the bolded lines showing moves recommended by sophisticated ex-ante deontology.

[16] Also known as the method of *backwards induction*. For a more detailed treatment of sophisticated choice, see Hammond (1988), McClennen (1990), Rabinowicz (1995), Buchak (2013: 170-200). The method itself appears to be due to Strotz (1955).

[17] Compare Frick (2015: 205) whose brand of contractualism includes a structurally similar "decomposition test".

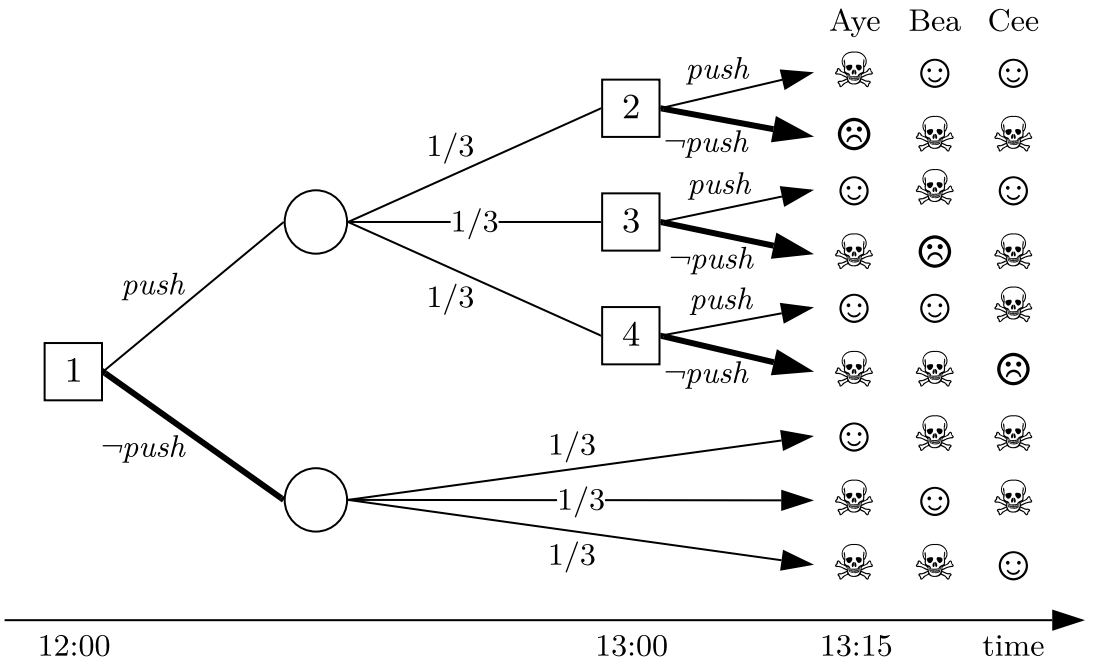


Figure 4-6. Case One with Sophisticated Choice

You know that, no matter what you find out at 13:00, you will then decide not to push the lever, since that will harm some people for the benefit of others while not being in everyone's interest. So, you know that deciding to push at 12:00 will be followed by not pushing at 13:00. Holding that fixed, your choice at 12:00 is really the following.

> *Push*: Everyone gets a 2/3 chance of death and a 1/3 chance of mild trauma.
>
> ¬*Push*: Everyone gets a 2/3 chance of death and a 1/3 chance of life.

Since the latter is in everyone's interest at 12:00, sophisticated ex-ante deontology tells you not to push at 12:00, letting the trolley run its course. Because of that, sophisticated ex-ante deontology does not conflict with sequential dominance nor dynamic consistency in Case One.

This is cold comfort, however, as it still has to give up sequential dominance in the barely different

> *Case Two.* Aye, Bea and Cee are trapped in three suitcases. The situation is the same as in Opaque Footbridge, except for the following details. It is now 12:00. You still don't know who is where but, luckily, everyone will peek out of their suitcases by 13:00. The trolley is slow and will only roll under the footbridge at 13:15. This time the lever is well-oiled but you are not allowed to leave before 13:00. You can either push the lever at 12:00, dropping the footbridge suitcase onto the track, or you can wait to make your decision until 13:00, just before you go. If you do decide to push the lever at 12:00, the two survivors on the track will be mildly traumatized, having to endure an extra hour of agonized screaming from whoever was on the footbridge. That won't happen if you instead decide to push at 13:00.

We can represent your decision problem as follows.

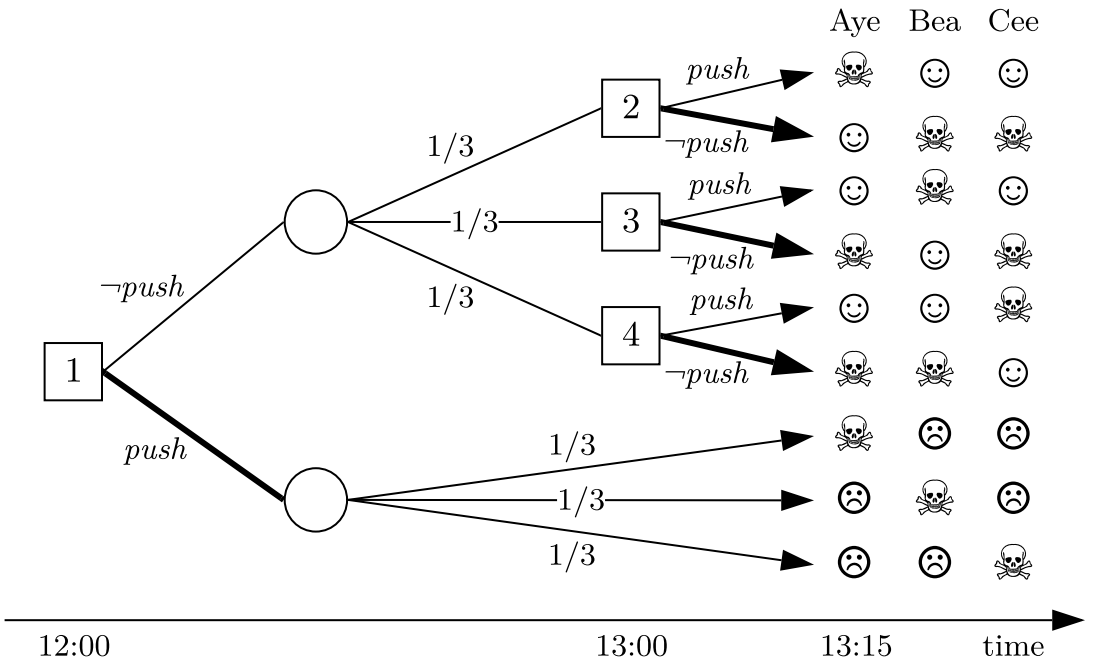


Figure 4-7. Case Two with Sophisticated Choice

You know you will not push the lever if you wait, since pushing will still be harming some to help others while no longer being in everyone's interest. So, you know that at 13:00 you would follow through on your 12:00 decision not to push. Holding that fixed, your choice at 12:00 is really the following.

> $\neg$*Push*: Everyone gets a 2/3 chance of death and a 1/3 chance of life.
>
> *Push*: Everyone gets a 1/3 chance of death and a 2/3 chance of mild trauma.

It is plausible that if the psychological trauma is mild enough, it is in everyone's interest that you push at 12:00. So, this is what sophisticated ex-ante deontology tells you to do.

But this means that you end up with one dead body and two cases of mild psychological trauma when you could have ended up with just one dead body and two happy survivors instead. This shows that sophisticated ex-ante deontology is

incompatible with sequential dominance, after all. This is both troubling in itself but also awkward given that sophisticated ex-ante deontology tries to preserve the idea that one should act in everyone's interest if possible.

Case Two also shows that sophisticated ex-ante deontology exhibits troubling information avoidance. After all, in Case Two, your choice at 12:00 is effectively between pushing the lever right away and waiting to find out who is where before you do. Sophisticated ex-ante deontology tells you to do the former, even though that means traumatizing two people in the process. So, sophisticated ex-ante deontology tells you to avoid relevant information at the expense of human happiness. This seems both irrational and immoral. We can put this point by saying that sophisticated ex-ante deontology violates

> **Good's Principle**. One should not avoid free information relevant to one's choice.[18]

While Good's principle is not uncontested, I think anyone rejecting it owes us some explanation. For example, Ahmed and Salow (2019) argue that Good's principle fails in cases where learning is risky, in the sense of potentially making some actual bad things appear unlikely. But why would learning be risky for a deontologist in Case Two?

What's more, sophisticated ex-ante deontology can make you wilfully rid yourself of relevant information that you already possess. This is *information disposal*, not just information avoidance.

To see this, consider the following variation on Cases One and Two. At 12:00 you know who is where, but you can take an amnesia pill that will make you lose your

---

[18] Compare Bradley and Steele (2016) and Buchak (2013: 191-200). The locus classicus is Good (1967).

knowledge. Whether or not you take it, you will later have to decide between pushing and not. Let's also assume that Aye, Bea and Cee will never find out who is where. Since they know that you follow sophisticated ex-ante deontology, they will be *begging* you to take the pill, at least if they are self-interested. They know that if you do take the pill, you will push the lever, and if you don't, you won't. So, they know that if you take the pill, they will get a 1/3 chance of death and a 2/3 chance of life, and if you don't, they will get a 2/3 chance of death and a 1/3 chance of life. Sophisticated ex-ante deontology therefore recommends the former. This is hardly plausible.

Similar problems beset sophisticated versions Hare's and Setiya's views. Let's consider Hare's view first. Combined with sophisticated choice, it arguably *requires* you to push at 12:00, hence violating both sequential dominance and Good's principle.

To see why, note that Hare permits you not to push if you know who is where. So, you know that if you wait at 12:00, then you *might* end up *not* pushing at 13:00. Holding that fixed, your choice at 12:00 is really the following.[19]

> ¬*Push*: Everyone *might* get a 2/3 chance of death and a 1/3 chance of life or everyone *might* get a 1/3 chance of death and a 2/3 chance of life.
>
> *Push*: Everyone is *sure* to get a 1/3 chance of death and a 2/3 chance of mild trauma.

Imagine that Aye, Bea and Cee know that you are sophisticated Hare follower, while still not knowing where they are. If they are self-interested, they will be *begging* you to push the lever right away. Pushing at 12:00 is definitely going to

[19] At least if we assume, plausibly enough, that whether you will push doesn't depend on who you will see on the footbridge. You give no special treatment to anyone.

bring their chance of death down to 1/3, at a small cost of mild trauma. By contrast, they will see waiting at 12:00 as pointlessly risky: it *might* give them prospects which are barely better than that, but it *might* also give them prospects which are markedly worse, by giving them a 2/3 chance of death. Since pushing at 12:00 is what everyone would ask for, if they were rational and self-interested, that's what a sophisticated Hare follower will do.[20]

Hare does try to pre-empt objections about information avoidance by saying that

> "in cases like this, where things of incommensurable value (like people's lives) are at stake, it makes perfect sense to resist the judgments of your better informed self" (2016: 456-7).

To see what Hare has in mind, consider the following tabular representation of Opaque Footbridge.

[20] Aye, Bea and Cee will be even more insistent if they find out that you randomize in the face of incommensurability: tossing a fair coin to decide between two incommensurable actions. This is because they will then see pushing at 12:00 as giving them a 1/3 chance of death and a 2/3 chance of mild trauma, and waiting at 12:00 as a 50/50 gamble between living and dying. Whether sophisticated choosers assume their later selves to randomize in the face of *indifference* is discussed by Rabinowicz (1995: 595; 1997: 289), whether they should do so in the face of *incommensurability* is less often discussed. See Broome's (1992) review of Levi (1986).

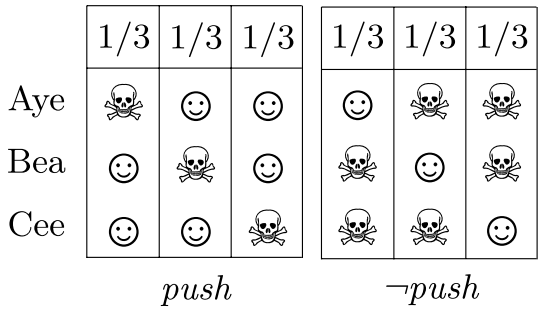


Figure 4-8. Opaque Footbridge in Tabular Form

The columns correspond to states of the world, specifying people's whereabouts. The rows correspond to people's prospects.

In that case, Hare thinks, pushing is mandatory, even though, in each state of the world, the outcome of pushing is incommensurable with the outcome of not pushing. So, one's better-informed self would not necessarily give the right advice about what one's less-informed self should do.

Perhaps this shows that it makes sense to *fail to defer* to the judgments of one's better-informed self. But in Case Two a sophisticated follower of Hare will *actively resist* these judgments and *actively resist* becoming better-informed. That, I think, does not make sense.

Moving on to Setiya's view, we see that it does no better in Case Two: it *permits* you to push at 12:00, thus violating sequential dominance and Good's principle. To see why, recall that Setiya forbids pushing if you know who is where. So, if you are a sophisticated follower of Setiya, you know you will not push at 13:00. Holding that fixed, your choice at 12:00 is really the following.

> ¬*Push*: Everyone gets a 2/3 chance of death and a 1/3 chance of life.
>
> *Push*: Everyone is sure to get a 1/3 chance of death and a 2/3 chance of mild trauma.

But this is pretty much the same choice as in Opaque Footbridge, where, recall, Setiya says that both pushing and not pushing is permissible. So, presumably, a sophisticated follower of Setiya might go either way in Case Two, thus pointlessly avoiding information and making some people gratuitously worse-off.

4 Resolute Ex-Ante Deontology

We saw that ex-ante deontology can tell us to pointlessly and predictably go back on our earlier decisions, and sometimes makes us gratuitously traumatize innocent people. A sophisticated implementation of ex-ante deontology does barely better: while it does not license fickleness, it can still license gratuitously traumatizing people. In addition, it can lead to irrational information aversion and information disposal.

Maybe the problem is not that the agent does not take into account what they are going to do later, but rather that they don't let their later actions depend on earlier history? This is the key idea behind the method of *resolute choice*.[21] We can revise ex-ante deontology accordingly:

> **Resolute Ex-Ante Deontology**. Continue a course of action that does or *did* give everyone best prospects, even if, as a result, some people will be harmed to help others. Don't harm people in that way otherwise.

Here is how this works in Case Two.

[21] See Machina (1989), McClennen (1990), Bader (2019), but also Buchak (2013: 170-200) and Rabinowicz (1995).

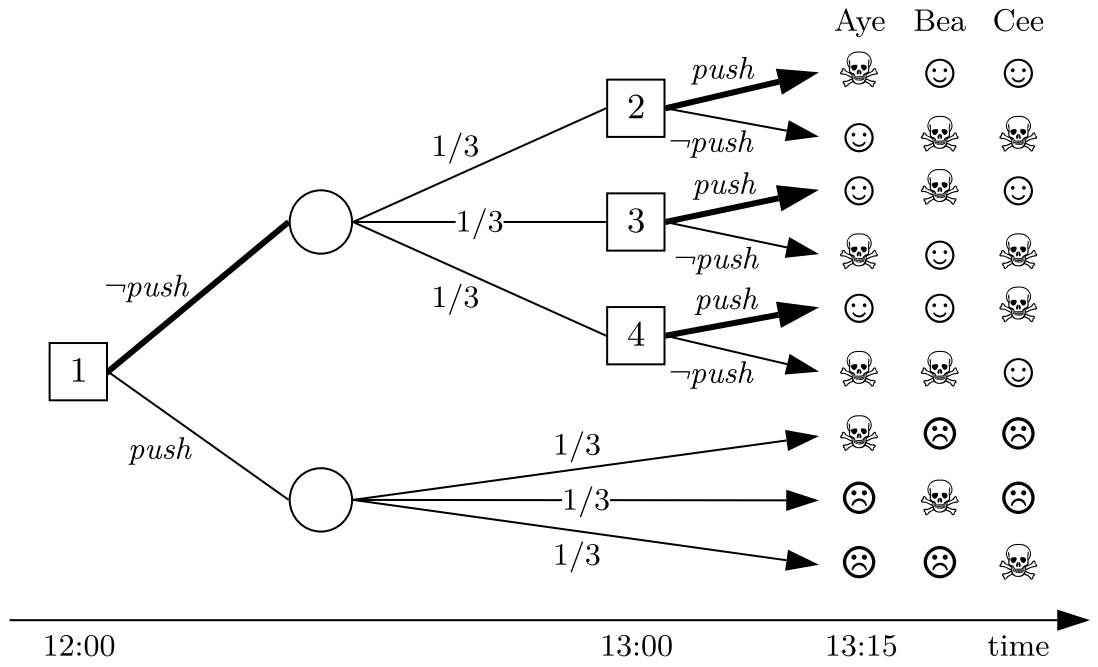


Figure 4-9. Case Two with Resolute Choice

At 12:00 everyone's interests are best served if you wait until 13:00 and push then. This means going up at node 1.

At 13:00 pushing is no longer in everyone's interest. Still, according to resolute ex-ante deontology, you should stick to your earlier plan, as pushing at 13:00 *was* in everyone's interest at 12:00. This means going up at nodes 3, 4, 5.

It is easily seen that, unlike ex-ante deontology and sophisticated ex-ante deontology, *resolute* ex-ante deontology does not run into problems with dynamic consistency, sequential dominance, nor Good's principle.

The view also had a prominent defender: Thomson (1990: 176-202).[22] She would say that pushing is mandatory in Opaque Footbridge, but only as an exception to a general prohibition on killing:

[22] Her position seems to have changed by the time of Thomson (2008).

> "The exceptions (…) are those in which the one who will be killed, and the [two] who will be saved, are members of a group such that it was to the advantage of all the members that the one (whoever he or she would later turn out to be) would later be killed, and the only thing that has since changed is that it is now clear who the one was going to turn out to be" (195).[23]

Thomson therefore thinks that pushing the lever is forbidden in Transparent Footbridge. In that case, there is no preceding time at which it is in everyone's interest to push the lever. Unfortunately, resolute ex-ante deontology faces some serious problems.

The first is simply that it makes the permissibility of harming some to help others depend on seemingly irrelevant facts about the past.

To see this, imagine it is now 2020 and you face a case like Transparent Footbridge. You can see who is in which suitcase. This is no surprise to you. In 1990 you foretold that in 2020 you will find yourself in a Transparent Footbridge case involving Aye, Bea and Cee. Does it matter for what you do now whether you simultaneously foretold who would be in which suitcase? Resolute ex-ante deontology says that it does. But that is very hard to believe.

As this example also shows, resolute ex-ante deontology can tell us to go *very far* into the past to determine what is to be done now. But how far exactly? Can we ever stop? And why stop at any particular place? This is the second problem.[24]

---

[23] Thomson uses "advantage" in a somewhat technical sense, see (1990: 184), but that nicety need not distract us here.

[24] Similar points are made about the method of resolute choice in other contexts by Hammond (1983: 183), Machina (1989: 1651-1653), Gustafsson (2015: 1599), and Gustafsson (2018: 602).

The third and last problem is that resolute ex-ante deontology undermines case-based intuitions that drive many people to deontology in the first place. We can no longer be sure, for example, whether pushing is forbidden if we find ourselves in a case which *looks like* Transparent Footbridge. After all, permissibility might depend on the distant past! Firm intuitions about impermissibility of harming are no longer generally reliable.[25]

To avoid all these problems, deontologists might want to reject resolute ex-ante deontology and instead accept

> **Irrelevance of Past Ignorance**. If it is permissible to impose costs on some people for the sake of benefits to some other people, then it is permissible to do so whether or not the make-up of either group has been known in advance.[26]

I will end this section by noting that resolute implementations of Hare's and Setiya's views face the same problems as resolute ex-ante deontology. The resolute version of Hare's view *requires* pushing the lever in full knowledge of who is where, provided that people's whereabouts weren't always known, but *merely permits* it otherwise. The resolute version of Setiya's view, on the other hand, *permits* pushing the lever in the former case but *forbids* it otherwise. So, they both have to face up to the problems of resolute ex-ante deontology.

---

[25] Thoma (2019: 249) makes a related objection against a resolute implementation of Buchak's (2013) risk-weighted expected-utility theory.

[26] I want this condition to be a *weak* version of McClennen's (1990: 120-22) *separability* condition which, roughly, says that the past doesn't matter as long as the available outcomes of one's current action remain the same.

5 The Veil-of-Ignorance Argument

I conclude that resolute ex-ante deontology does not help where both ex-ante deontology and sophisticated ex-ante deontology fail. The reader might nonetheless wonder, justifiably, whether my arguments hit all the possible views spanning the space between permissivism and orthodox deontology.

I think they do. In this section I will present a recipe to rule out any views of this sort. The ingredients will include many of the theoretical desiderata I appealed to in my discussion above.

Let's start with the following simple case:

> *Case Three.* Two people are involved. It is now 12:00. A fair coin will be tossed at 13:00. After you see how it came up, you'll be able to intervene. If the coin comes up Heads and you intervene, then, at 13:15, Bea will be harmed to degree y for the sake of benefitting Aye to degree x. If the coin comes up Tails and you intervene, then, at 13:15, Aye will be harmed to degree y for the sake of benefitting Bea to degree x. If you don't intervene, neither will be harmed nor benefitted, regardless of how the coin comes up.

Your decision problem can be represented as follows.

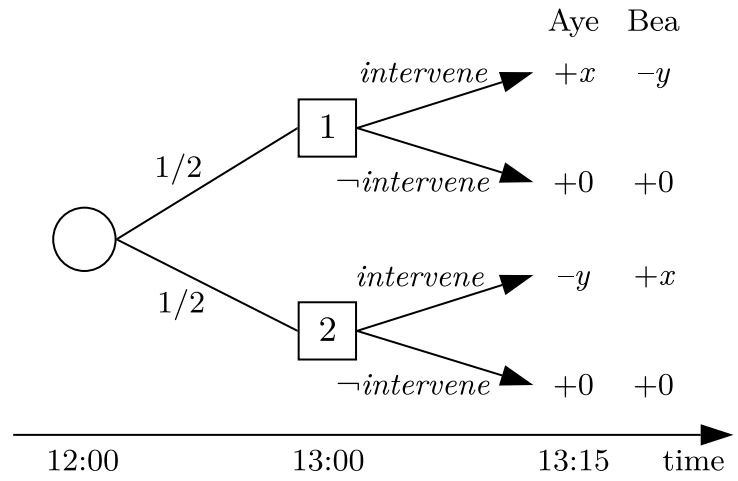


Figure 4-10. The Veil-of-Ignorance Argument

There are four courses of action open to you in this case. You can intervene no matter what, you can decline to intervene no matter what, or you can make your intervention conditional on the coin's outcome (intervening on Heads only or on Tails only). All in all, you can affect Aye's and Bea's prospects in the following way.

Intervene on Heads?	Intervene on Tails?	Aye	Bea
✔	✔	$+1/2x\,{-}1/2y$	$+1/2x\,{-}1/2y$
✔	×	$+1/2x$	$-1/2y$
×	✔	$-1/2y$	$+1/2x$
×	×	$+0$	$+0$

Figure 4-11. Possible actions in Case Three

What is the right course of action from the vantage point of 12:00?

First of all, we can rule out the middle two courses of action as discriminatory. Suppose either of them is permissible at 12:00. Then, by **dynamic consistency**, they would be permissible to continue at 13:00. But that would contradict

> **Symmetry**. Whether it is permissible to impose costs on some people for the sake of benefits to other people does not depend on people's identities but at most on the sizes of the groups and the magnitude of costs and benefits.

The only two remaining courses of action are intervening twice and never. Now suppose, for the sake of argument, that a 50/50 gamble between gaining x and losing y is in both Aye's and Bea's interest. If rational self-interest calls for maximizing expected gain, then it is enough to assume that x is greater than y.[27] Then at 12:00 everyone's interests are better served if you intervene twice than if you never intervene.

That you can only do the former follows from

> **Ex-Ante Pareto**. If it is in everyone's interest that some action be performed rather than another, then the latter cannot be permissible.

So, you are now left with intervening twice as the only remaining course of action. Assuming that this is no moral dilemma, intervening twice is therefore mandatory from the vantage point of 12:00. But, then, by **dynamic consistency**, you should follow through at 13:00 and intervene no matter how the coin lands.

Lastly, by the **irrelevance of past ignorance**, it follows that you should intervene even if there is no preceding uncertainty about the loser's identity, as in the following two cases.

---

[27] In general, the values of x and y to use here will depend on the rational attitude to risk: whether we should be risk-averse or risk-loving.

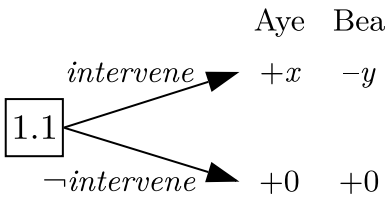


Figure 4-12

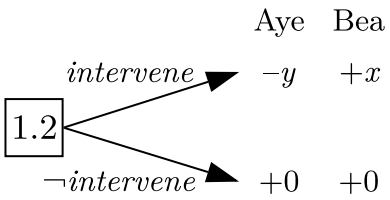


Figure 4-13

So, we can now conclude that if a 50/50 gamble between gaining x and losing y is in everyone's interest, then it is also mandatory to impose a loss of y on someone for the sake of bestowing a benefit of x on someone else.

By using a many-sided fair die instead of a two-sided fair coin, we can easily generalize the argument just given to establish

> **The Veil-of-Ignorance Principle**. If it is in everyone's interest to face a prospect of getting x_1 with probability $\frac{1}{n}$, x_2 with probability $\frac{1}{n}$, …, and x_n with probability $\frac{1}{n}$ rather than a prospect of y_1 with probability $\frac{1}{n}$, y_2 with probability $\frac{1}{n}$, …, and y_n with probability $\frac{1}{n}$, then it is mandatory to bring about an n-person people get $x_1, x_2, \dots, x_n$ rather than a distribution where they get $y_1, y_2, \dots, y_n$.[28]

So, we see that dynamic consistency, symmetry, and irrelevance of past ignorance can be used to connect individual choice between prospects with choice between allocations of benefits and harms, and to rule out deontic constraints on harming in the process.

---

[28] A similar principle, albeit in an axiological formulation, is discussed by Thomas (2016: 101). My overall argument is inspired by Thomas's and Hammond's (1996) arguments. See also Buchak (2017) and Nebel (2020).

The derivation just sketched therefore matters beyond the context of deontology. In particular, we can use it to bolster Harsanyi's (1953, 1977) and Vickrey's (1945) appeals to veils of ignorance by answering an important objection which Barry (1989) puts as follows:

> "Suppose I am initially disinclined to believe I must always be prepared to sacrifice my own interests whenever by doing so I can provide somebody else with a larger benefit, or whenever I can provide some larger number of people with small benefits that are cumulatively larger. I do not see that I have been given any adequate cause for changing my mind if I am told that, as a utility-maximizer behind a thin veil of ignorance, this is what I would have endorsed as a principle. I may agree that I would indeed have done so but then ask: 'So what?'" (334-335).[29]

My response to Barry's 'so what?' is twofold. First: veils of ignorance need not be *fictitious*, but can be *natural*.[30] Cases like Opaque Footbridge are cases of natural uncertainty and are much more realistic than Harsanyi's disembodied spectators choosing between societies. Second: principles of rational and moral choice in *sequential decision problems* can be used to connect choices behind such natural veils with actual unveiled moral choices. I take this as a win, even if a partial one, for Harsanyi's and Vickrey's approach to ethics.

---

[29] As Rabinowicz (2009) says in a related context: "pretence in, pretence out. With premisses we only pretend to accept, the conclusion wouldn't be accepted for real". See also Scanlon (1982), Broome (1991: 51-59), Thomas (2016: 99-102).

[30] Naturalness of some veils of ignorance is also noted by Frick (2015: 190), although he arguably would not endorse the use I make of it here.

6 Push in Opaque Footbridge?

So far, I have argued against compromise positions between permissivism and orthodox deontology. If I am right, the key issue is whether you should push in cases like Opaque Footbridge. If 'Yes', then we should accept permissivism, rejecting deontic constraints. If 'No', we are free to accept orthodox deontology instead.

I will now give a version of the most compelling argument for pushing that I know of. It starts from the idea that a person's interests are morally dispositive in cases where only they are affected.[31] If I am right, deontologists have to reject that idea. That, I think, is a serious drawback of their view.

The argument has two parts. The first is about the following case.[32]

> *Two Opaque Tracks.* There are now two tracks and two out-of-control lethal trolleys. Both tracks look exactly the same. No one knows who is where. Aye, Bea and Cee are in three suitcases around the first track. Two of them are on the track itself, in the trolley's path. The other is on the footbridge above the tracks. The three suitcases around the second track are filled with sand. You can swap suitcases from the first track with the corresponding suitcases on the second track, but without seeing where they come from or where they go. You have three buttons. Button 1 moves Aye's suitcase. Button 2 moves Aye's *and Bea's* suitcases. Button 3 moves Aye's, Bea's *and Cee's* suitcases. The catch is that by moving *any* of the suitcases, you will topple the footbridge suitcase on the second track. That

[31] Compare Hammond (1996), Crisp (2011), Frick (2015: 186-194), and Hare (2016: 455-464).

[32] This is a slight variant of Hare's (2016: 461-462) case.

suitcase will then be enough to halt the second trolley, whether it contains a person or is filled with sand.

What should you do? We can represent your decision problem as follows, where "p☠$+q$☺" denotes a prospect carrying a p chance of dying and a q chance of living.

	Aye	Bea	Cee
Button 3	1/3☠+2/3 ☺	1/3☠+2/3 ☺	1/3☠+2/3 ☺
Button 2	1/3☠+2/3 ☺	1/3☠+2/3 ☺	2/3☠+1/3 ☺
Button 1	1/3☠+2/3 ☺	2/3☠+1/3 ☺	2/3☠+1/3 ☺
Do nothing	2/3☠+1/3 ☺	2/3☠+1/3 ☺	2/3☠+1/3 ☺

Figure 4-14. Two Opaque Tracks

The decision to move each extra suitcase affects only one individual. Pressing button 1 only affects Aye, improving her lot from a 2/3 chance of death to a 1/3 chance of death. Switching from button 1 to button 2 only affects Bea, again improving her lot in the same way. Likewise for Cee's and the switch from button 2 to 3. So, intuitively, you should press button 3. Why?

Suppose you decide to do nothing. Then Aye can rightly complain: "Hey! You could at least move my suitcase. That will benefit me and not affect anyone else". It seems that Aye has a point: you should move at least *one* suitcase. So, you decide to press button 1. But then Bea can make the same complaint as Aye just did, while also accusing you of arbitrariness ("Why listen to Aye and not to me?!"). Likewise if you decide to press button 2. It is only by pressing button 3 that you escape any complaints. So, it seems, you should press it.

Let's consider two objections at this point. First, following Setiya, someone might object that "there are powerful reasons not to intervene in others' lives even when they would benefit from intervention" (2020: 65), namely, reasons to respect

people's autonomy. This suggests that moving *any* suitcase is either wrong or, at least, not mandatory. I think this is misguided. If I save you from drowning in a shallow pond, do I thereby diminish your autonomy? Clearly, I don't, unless you jumped in there to commit suicide or to make an artistic statement or the like. And we can assume that Aye, Bea and Cee harbour no death wishes of that sort. So, while reasons of autonomy might be real enough, they are irrelevant here.

The second objection is that deontologists have a principled reason to resist pressing button 3. To see why, note that pressing button 1 carries zero chance of using some people for the benefit of others, pressing button 2 increases that chance to 2/3, while pressing button 3 takes it up to unity. Deontologists might say that the chance of 2/3 is unacceptably high, deciding to stop at button 1 instead. I think that would reveal unpalatable *causal fetishism* on their part: they would have to care more about the causal structure of the situation than about the people affected.[33] After all, both Aye and Bea are happy to take an extra chance of being used as a train stop for the sake of a greater chance of survival. So, deontologists cannot really be worried about the only two people at stake when they resist pressing button 2. They must be worried about impersonal causal structure instead.[34]

---

[33] Compare McMahan: "it is difficult to believe that the way in which an agent is instrumental in the occurrence of an outcome could be more important than the nature of the outcome itself. (…) Is it really credible to suppose that how one acts on that single occasion matters more in moral terms than the whole of the life that will be lost if one lets the two die rather than killing the one?" (1993: 279). See also Norcross (2008).

[34] Hare (2013: 89-96) has another argument for pushing in Opaque Footbridge that *starts* from the related idea that one shouldn't obsess with one's *dirty hands*. My idea of causal fetishism is less agent-centric and comes at a different place in my argument.

I conclude that even deontologists should find it hard to resist the case for moving all three suitcases in Two Opaque Tracks. But why is Two Opaque Tracks even relevant to Opaque Footbridge where you cannot affect Aye, Bea, and Cee individually? This is where the second part of the argument comes in, and it is simply that it would be implausible to endorse moving all three suitcases in Two Opaque Tracks but to endorse *not* pushing the lever in Opaque Footbridge. That would be like saying: "Well, if only I could do what's best for everyone, one person at a time! Then, of course, I would. But since I can only do what's best for everyone *en bloc* or not at all, I won't!" It seems that adding or removing these extra choices shouldn't make a difference.[35]

So, those who want to avoid pushing in Opaque Footbridge either have to say that adding some extra innocent choices makes a difference in permissibility or that sometimes it is right to ignore complaints of the only person with anything at stake. Both options are unappealing, the latter likely unpalatably fetishistic. I conclude that you should push in Opaque Footbridge. But, if I am right about ex-ante deontology and other compromise views like it, this means you should push in Transparent Footbridge, too. This, arguably, means giving up on deontology. Deontologists have to get off the train to that conclusion. But where?[36]

---

[35] We could here appeal to a version of Sen's (1969) property α. Compare Rulli and Worsnip (2016), but also Sen (1993: 500) and Temkin (2012: 387-390).

[36] The reader might now wonder whether my argument is any better than Hare's who also appeals to a case like Opaque Two Tracks. There are two main differences: my argument is nonsequential and it does not appeal to principles of deontic logic like Hare's own principle of *ought agglomeration*. It is also worth noting that Hare's ought agglomeration principle can be shown to imply acyclicity of moral choiceworthiness in cases structurally analogous to Rabinowicz's (2000) persistent money pump.

7 Problems for Minimally Paretian Deontology

Even if deontologists cannot always act in everyone's interest in single-person choices, we might think that this only happens in cases where there is uncertainty about the true state of the world. So, perhaps, they can still accept

> **Minimally Paretian Deontology**. Choose an action that gives everyone best *outcomes*, even if, as a result, some people are harmed to help others. Don't harm people in that way otherwise.

This is much more limited than any form of ex-ante deontology I considered above and it does not face any of its problems, as it says nothing about cases where there is uncertainty about the true state of the world. It can also be motivated by the desire to avoid causal fetishism. If using some to help others happens to actually make everyone better-off, then, other things being equal, what could possibly be wrong with it?

Still, even this non-fetishistic form of deontology has a serious problem. We can see it in the following variation of Opaque Footbridge:

> *Case Four.* Aye, Bea and Cee are trapped in three suitcases. The situation is the same as in Transparent Footbridge, except for the following details. You now have the power to decide *whether* the footbridge suitcase will be toppled onto the track, halting the trolley from hitting the other two suitcases, but also to decide *who* will be on the footbridge when this happens, by switching around the relevant suitcases. You can do this by pressing any of the following six buttons.
>
> > *Button 1.* Put Aye on the footbridge, do *not* topple the suitcase.
> >
> > *Button 2.* Put Bea on the footbridge, topple the suitcase.
> >
> > *Button 3.* Put Bea on the footbridge, do *not* topple the suitcase.

Button 4. Put Cee on the footbridge, topple the suitcase.

Button 5. Put Cee on the footbridge, do *not* topple the suitcase.

Button 6. Put Aye on the footbridge, topple the suitcase.

The buttons are wired together, so that only one can remain pressed at any one time. Any pressing will only take effect in a few minutes, so you can change your mind if you switch fast enough. When you come onto the scene, you see that button 1 is already pressed.

We can represent your decision problem as follows.

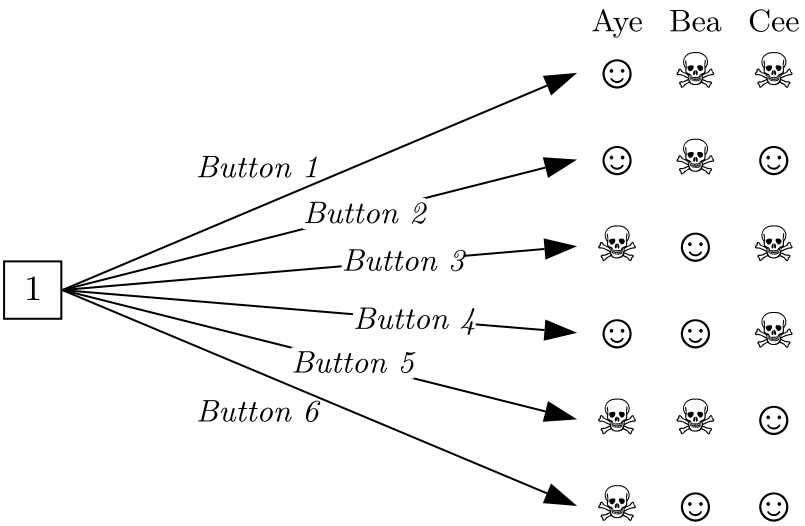


Figure 4-15. Case Four

What should you do? The problem is that minimally Paretian deontology seems to rule out any possible answer to that question, as it here leads to *deontic cycling*.[37] To see how, consider the following train of thought.

[37] A similar, albeit distinct, cycle is discussed by Willenken (2012). I take my cycle to be much more troubling than other deontic cycles considered by Kamm (1996: 311-354) and Temkin (2012: 194-231).

We start out with Aye on the footbridge. Aye is going to live, Bea and Cee are going to die. You are about to walk away. At that point Cee could complain: "Hey! Why let that happen to me? Why not put Bea on the footbridge and topple the suitcase? That won't make a difference to Aye (who will be fine either way) and Bea (who will be dead either way). But that will make a huge difference to *me*: it will save my life!" You hear what Cee is saying and, not wanting to be a causal fetishist, you press button 2.

But then a deontologist standing by interjects: "Hey! If you settle on that, you will be simply using Bea for the benefit of Aye and Cee! That's wrong". As a good deontologist yourself, you heed their advice and press button 3 instead.

At that point Aye could complain: "Hey! Why do that to me? Why not put Cee on the footbridge and topple the suitcase?" You listen to them, and quickly decide to press button 4. And so on: switching between buttons until the time is up. It seems that there is no stable resting point for minimally Paretian deontologists to occupy. The case of deontic cycling at hand appears vicious.[38]

I think this is a strike against deontology. If you listen to the stand-by deontologist in Case Four, you will become a causal fetishist, ignoring the interests of the only person with anything at stake for the sake of impersonal causal structure. Deontologists cannot be even minimally Paretian. They have to be causal fetishists even in cases where there is no uncertainty about the true state of the world.

[38] There are benign deontic cycles where it is relatively clear what should be done when all links of the cycle are available at once. See, for example, the Condorcet-style cycle in Gustafsson (2015).

8 Conclusion

What have I established? I first argued that deontologists cannot make room for the appealing idea that acting in everyone's interest is sometimes mandatory, even if it ensures that some people are harmed for the benefit of others. This was the upshot of my argument against the three forms of ex-ante deontology. I then gave an argument for that appealing idea. Rejecting it commits deontologists to unpalatable causal fetishism. Then I showed why deontologists cannot act in everyone's interests, even in cases where there is no uncertainty about the true state of the world. All in all, it appears that, contrary to what many deontologists have been saying for years, the only way to respect flesh-and-blood persons rather than the abstract idea of personhood is to embrace a doctrine closer in spirit to consequentialism, rejecting deontic constraints.

9 References

Adler, M. (2012). *Well-being and fair distribution*. Oxford: Oxford University Press.

Adler, M. D., & Sanchirico, C. W. (2006). Inequality and uncertainty: Theory and legal applications. *University of Pennsylvania Law Review, 155*(2), 279-377. doi:10.2307/40041309

Ahmed, A., & Salow, B. (2019). Don't look now. *The British Journal for the Philosophy of Science, 70*(2), 327-350. doi:10.1093/bjps/axx047

Alexander, L. (2014). The means principle. Unpublished manuscript. https://ssrn.com/abstract=2378608

Bader, R. (2019). Agent-relative prerogatives and sub-optimal beneficence. *Oxford Studies in Normative Ethics 9*.

Barry, B. (1989). *Theories of justice*. Berkeley: University of California Press.

Bradley, S., & Steele, K. (2016). Can free evidence be bad? value of information for the imprecise probabilist. *Philosophy of Science, 83*(1), 1-28. doi:10.1086/684184

Broome, J. (1991). *Weighing goods: Equality, uncertainty and time*. Oxford: Basil Blackwell.

Broome, J. (1992). Hard choices: Decision making under unresolved conflict by Isaac Levi. (review). *Economics and Philosophy, 8*(1), 169-176. doi:10.1017/S0266267100000560

Buchak, L. (2013). *Risk and rationality*. Oxford: Oxford University Press.

Buchak, L. (2017). Taking risks behind the veil of ignorance. *Ethics, 127*(3), 610-644. doi:10.1086/690070

Crisp, R. (2011). In defence of the priority view: A response to Otsuka and Voorhoeve. *Utilitas, 23*(1), 105-108. doi:10.1017/S0953820810000488

Frick, J. (2015). Contractualism and social risk. *Philosophy & Public Affairs, 43*(3), 175-223. doi:10.1111/papa.12058

Fried, B. H. (2012). What does matter? the case for killing the trolley problem (or letting it die. *Philosophical Quarterly, 62*(248), 505-529. doi:10.1111/j.1467-9213.2012.00061.x

Glover, J. (1977). *Causing death and saving lives*. Harmondsworth: Penguin.

Good, I. J. (1967). On the principle of total evidence. *The British Journal for the Philosophy of Science, 17*(4), 319-321. doi:10.1093/bjps/17.4.319

Gustafsson, J. E. (2015). Sequential dominance and the anti-aggregation principle. *Philosophical Studies, 172*(6), 1593-1601. doi:10.1007/s11098-014-0366-0

Gustafsson, J. E. (2018). The difference principle would not be chosen behind the veil of ignorance. *Journal of Philosophy, 115*(11), 588-604. doi:10.5840/jphil20181151134

Gustafsson, J. E., & Espinoza, N. (2010). Conflicting reasons in the small-improvement argument. *Philosophical Quarterly, 60*(241), 754-763. doi:10.1111/j.1467-9213.2009.648.x

Hammond, P. (1983). Ex-post optimality as a dynamically consistent objective for collective choice under uncertainty. In P. K. Pattanaik, & M. Salles (Eds.), *Contributions to economic analysis* (pp. 175-205) Elsevier.

Hammond, P. (1988). Consequentialist foundations for expected utility. *Theory and Decision, 25*(1), 25-78. doi:10.1007/BF00129168

Hammond, P. (1996). Consequentialist decision theory and utilitarian ethics. In F. Farina, F. Hahn & S. Vannucci (Eds.), *Ethics, rationality, and economic behaviour* (pp. 92-118). Oxford: Oxford University Press.

Hare, C. (2016). Should we wish well to all? *Philosophical Review, 125*(4), 451-472.

Hare, C. J. (2013). *The limits of kindness*. Oxford: Oxford University Press.

Harris, J. (1975). The survival lottery. *Philosophy, 50*(191), 81-87. doi:10.1017/S0031819100059118

Harsanyi, J. (1977). Morality and the theory of rational behavior. *Social Research, 44*(4), 623.

Harsanyi, J. C. (1953). Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy, 61*(5), 434-435. doi:10.1086/257416

Kagan, S. (1989). *The limits of morality* Oxford University Press.

Kamm, F. (1993). Non-consequentialism, the person as an end-in-itself, and the significance of status. *Philosophy and Public Affairs, 22*(1), 354.

Kamm, F. M. (1996). *Morality, mortality: Volume II: Rights, duties, and status*. Oxford: Oxford University Press.

Kamm, F. M. (2007). *Intricate ethics* Oxford University Press.

Levi, I. (1986). *Hard choices: Decision making under unresolved conflict*. Cambridge: Cambridge University Press.

Machina, M. J. (1989). Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature, 27*(4), 1622-1668.

McClennen, E. F. (1990). *Rationality and dynamic choice: Foundational explorations*. Cambridge: Cambridge University Press.

McMahan, J. (1993). Killing, letting die, and withdrawing aid. *Ethics, 103*(2), 250-279. doi:10.1086/293495

Nebel, J. M. (forthcoming). Rank-weighted utilitarianism and the veil of ignorance. *Ethics*. https://philpapers.org/archive/NEBRUA.pdf

Norcross, A. (2008). Off her trolley? Frances Kamm and the metaphysics of morality. *Utilitas, 20*(1), 65-80.

Nozick, R. (1974). *Anarchy, state, and utopia*. Basic Books.

Otsuka, M. (2011). Are deontological constraints irrational? In R. Bader, & J. Meadowcroft (Eds.), *The Cambridge companion to Nozick's Anarchy, State, and Utopia* (pp. 38-58) Cambridge University Press.

Peterson, M. (2015). Prospectism and the weak money pump argument. *Theory and Decision, 78*(3), 451-456. doi:10.1007/s11238-014-9435-2

Rabinowicz, W. (1995). To have one's cake and eat it, too: Sequential choice and expected-utility violations. *The Journal of Philosophy, 92*(11), 586-620. doi:10.2307/2941089

Rabinowicz, W. (1997). On Seidenfeld's criticism of sophisticated violations of the independence axiom. *Theory and Decision, 43*(3), 279-292.

Rabinowicz, W. (2000). Money pump with foresight. In M. Almeida (Ed.), *Imperceptible harms and benefits. Library of ethics and applied philosophy, vol. 8.* Springer, Dordrecht.

Rabinowicz, W. (2009). Preference utilitarianism by way of preference change? In T. Grüne-Yanoff, & S. O. Hansson (Eds.), *Preference change: Approaches from philosophy, economics and psychology* (pp. 185-206). Dordrecht: Springer Netherlands. doi:10.1007/978-90-481-2593-7_9

Scanlon, T. M. (1982). Contractualism and utilitarianism. In A. Sen, & B. Williams (Eds.), *Utilitarianism and beyond* (pp. 103-128). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511611964.007

Scheffler, S. (1994 [1982]). *The rejection of consequentialism: A philosophical investigation of the considerations underlying rival moral conceptions* (Rev. ed.). Oxford: Clarendon Press.

Sen, A. (1969). Quasi-transitivity, rational choice and collective decisions. *The Review of Economic Studies, 36*(3), 381-393. doi:10.2307/2296434

Sen, A. (1993). Internal consistency of choice. *Econometrica, 61*(3), 495-521. doi:10.2307/2951715

Setiya, K. (2020). Ignorance, beneficence, and rights. *Journal of Moral Philosophy, 17*(1), 56-74. doi:10.1163/17455243-20182841

Singer, P. (1977). Utility and the survival lottery. *Philosophy, 52*(200), 218-222. doi:10.1017/S0031819100023172

Strotz, R. H. (1955). Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies, 23*(3), 165-180. doi:10.2307/2295722

Temkin, L. S. (2012). *Rethinking the good: Moral ideals and the nature of practical reasoning.* New York; Oxford: Oxford University Press.

Thoma, J. (2018). Risk aversion and the long run. *Ethics, 129*(2), 230-253. doi:10.1086/699256

Thomas, T. (2016). Topics in population ethics. DPhil thesis. University of Oxford. https://ora.ox.ac.uk/objects/uuid:fa2a09aa-e784-4126-bd4a-0487d3653add

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal, 94*(6), 1395-1415. doi:10.2307/796133

Thomson, J. J. (1990). *The realm of rights*. Cambridge, Mass; London: Harvard University Press.

Thomson, J. J. (2008). Turning the trolley. *Philosophy & Public Affairs, 36*(4), 359-374. doi:10.1111/j.1088-4963.2008.00144.x

Rulli, T., and Worsnip, A. (2016). IIA, rationality, and the individuation of options. *Philosophical Studies, 173*(1), 205-221. doi:10.1007/s11098-015-0481-6

Vickrey, W. (1945). Measuring marginal utility by reactions to risk. *Econometrica, 13*(4), 319-333. doi:10.2307/1906925

Willenken, T. (2012). Deontic cycling and the structure of commonsense morality. *Ethics, 122*(3), 545-561. doi:10.1086/664750

Chapter 5

Transfinitely Transitive Value

Abstract: This paper develops transfinite extensions of transitivity and acyclicity, and applies them to the problem of evaluating outcomes where different numbers of people exist. Transfinite transitivity is used to argue that it is better to add good lives, worse to add bad lives, and equally good to add neutral lives, where a life's value is understood as personal value. These conclusions rule out a number of theories of population ethics, feed into an argument for the repugnant conclusion, and allow us to reduce different-number comparisons to same-number ones. Challenges to these arguments are addressed, including the issue of comparing existence and nonexistence in terms of personal value, the possibility of minimal quanta of time and life, and the meaningfulness of measuring closeness between outcomes with different population sizes. An asymmetry is uncovered between transfinite cycles of worseness and betterness, leading to a new justification for a weak form of procreative asymmetry. Transfinite transitivity and related principles are also favourably compared to the better-known principles of continuity. Transfinite transitivity and acyclicity principles therefore promise to break new ground in population ethics.[1]

Word count: 9597

[1] I would like to thank Michal Masny, Todd Karhu, Aidan Penn, Jack Kelly, and Tushar Menon. Special thanks to Ralf Bader, Teru Thomas, Theron Pummer, and Tomi Francis who gave me helpful written comments on an earlier draft.

1 Introduction

According to

> **Transitivity of Strict Betterness**. For all value-bearers X, Y, Z, if X is better than Y and Y is better than Z, then X is better than Z.[1]

Here is one way to appreciate the meaning of this principle. Take a sequence $X_1, X_2, ...$ such that X_1 is better than X_2, which in turn is better than X_3, which … and so on. If X comes at the end of such a sequence, transitivity allows us to make a shortcut: we can conclude that X_1 is better than X. Provided that the sequence is finite, that is. *Transfinite* transitivity of betterness relaxes this finiteness proviso, allowing us to make infinite as well as finite shortcuts. What this means exactly will be explained in the next section.

Transitivity is a powerful principle, as evidenced by Parfit's mere addition paradox (1987 [1984]: chapter 19), Broome's argument against the neutrality intuition (2004: chapter 10) and Temkin's (1996) spectra arguments. Rachels (2001) and Temkin (1996, 2012) take these unexpected results to cast doubt on transitivity itself.

Transfinite transitivity is much more powerful than that. This paper will show how to use it to argue that creating people with good lives is better than not creating them and that creating people with bad lives is worse than not creating them, thus significantly narrowing down the range of possible views in population ethics.

[1] "Value-bearers" are all and only things in the field of the "at least as good" relation, that is, things which are better, worse, or just as good as something.

Why believe transfinite transitivity? In my other work, here included as Chapter 6, I offer a conditional argument that if transitivity is defensible, so is transfinite transitivity.[2] Its central part concerns transfinite transitivity's role in theory of rational choice. In an important range of infinite decision problems, it can play the same role that transitivity plays in finite decision problems: it helps secure the existence of best options, it can be characterized by consistency conditions on permissible choice, and it can be supported by money-pump arguments. This suggests transfinite transitivity truly generalizes transitivity.

The next section, 2, explains transfinite transitivity and related principles in more detail. Sections 3 and 4 show how to use them in population ethics. Section 5 then takes up some objections and loose ends regarding these arguments. Section 6 is about how pragmatic (money-pump) arguments bear on the status of our transfinite principles in the context of population ethics. This discussion suggests a novel justification for a version of the so-called procreative asymmetry. Section 7 compares transfinite transitivity to a related but better-known principle of continuity, showing how the former is importantly more liberal.

[2] Since I agree with Broome (2004: 50-63) about transitivity's truth, I suggest we take this as a *modus ponens*. But even transitivity sceptics can agree with me on the relevant logical entailments and do a *modus tollens* instead.

2 Principles

Transitivity of strict betterness can be generalized into the transfinite whenever our domain of evaluation is equipped with a notion of *convergence* which is perhaps easiest to explain in terms of *closeness* (or *distance*). We say that the sequence $X_1, X_2, ...$ *converges* to X (its *limit*) iff it eventually becomes arbitrarily close to X. For example, the sequence of numbers $1, \frac{1}{2}, \frac{1}{4}, ...$ converges to number 0, since the absolute difference between members of that sequence and 0 is eventually arbitrarily small.[3] How to understand closeness in the case of value-bearers will depend on the application at issue, and we will see some examples below.

With convergence in hand, we can build towards *transfinite* transitivity of strict betterness by noting that ordinary transitivity allows us to make shortcuts along finite paths of strict betterness, as follows.[4]

$$X_1 \succ X_2 \succ X_3 \text{ implies } X_1 \succ X_3.$$

$$X_1 \succ X_2 \succ X_3 \succ X_4 \text{ implies } X_1 \succ X_4.$$

$$X_1 \succ X_2 \succ X_3 \succ X_4 \succ X_5 \text{ implies } X_1 \succ X_5.$$

...

[3] More precisely, for any positive number ϵ, all elements sufficiently far in the sequence are within distance ϵ of x. This assumes we are working with *metric spaces*, a subset of *topological spaces*, where some distance metric is defined. For more explanation of all mathematical concepts in this paper, see a standard topology textbook such as Mendelson (1962).

[4] Notation: "$\succcurlyeq$" will mean "at least as good as", "$\succ$" "better than", "$\prec$" "worse than", "$\preccurlyeq$" "at least as bad as", and "$\sim$" will mean "equally good as". I will call these relations "weak betterness", "strict betterness", "strict worseness", "weak worseness" and "equal goodness", respectively.

Here all $X_1, X_2, ...$ are value-bearers. Transfinite transitivity extends this pattern into the transfinite, by means of the notion of convergence: for all value-bearers $X, X_1, X_2, ...$,

$$X_1 \succ X_2 \succ X_3 \succ \cdots \to X \text{ implies } X_1 \succ X,$$

where "$\to$" indicates convergence.[5] Put differently:

> **Transfinite Transitivity of Strict Betterness**. For all value-bearers $X, X_1, X_2, ...$, if the sequence $X_1, X_2, ...$ converges to X, and X_1 is better than X_2, X_2 is better than X_3, …, then X_1 is better than X.

Analogously, we can define transfinite transitivity of other binary relations such as weak betterness, strict worseness, weak worseness, or equal goodness. Principles from the finite case that are weaker than transitivity can also be generalized into the transfinite. For example, we have

> **Acyclicity of Strict Betterness**. For all value-bearers $X_1, X_2, \ldots, X_n$, if X_1 better than X_2, X_2 is better than X_3, …, and X_{n-1} is better than X_n, then X_n is not better than X_1.

This generalizes to

> **Transfinite Acyclicity of Strict Betterness**. For all value-bearers $X, X_1, X_2, \ldots, X_n$, if the sequence $X_1, X_2, ...$ converges to X, and X_1 is better than X_2, X_2 is better than X_3, …, then X is not better than X_1.

---

[5] This notation is from Bartha, Barker & Hájek (2014) who also introduce the name "transfinite transitivity" for what I will call "transfinite acyclicity of strict worseness". The idea of transfinite transitivity itself is apparently due to Gillies (1959). It was then reinvented by Smith (1974) and discussed by Birchenhall (1977), Mukherji (1977) (where transfinite acyclicity of strict worseness in my sense is introduced) and later by Carosi & Zaffaroni (1990) and Kukushkin (2008). In philosophy it was reinvented again by both Bartha et al. (2014) and Weatherson (ms).

These definitions work for other binary relations, too. Transfinite transitivity of strict betterness expresses the thought that whenever things are getting worse, in the limit they must be the worst, while transfinite acyclicity expresses the logically weaker thought that whenever things are getting worse, they cannot be better in the limit.

To be clear, these transfinite principles differ from their finite namesakes in that they require some notion of convergence, here understood in terms of closeness (or distance). Whether the principles are appealing or even meaningful depends on that further choice. But, in many cases, we do have an intuitive grasp of convergence.

3 Arguments

What is the value of adding an extra life? In a sense that is the basic question of population ethics. Transfinite transitivity principles can help us answer it.

3.1 Constant additions

First consider adding a life that is *constantly neutral*, in the sense that it is equally good no matter how long it is. If we graphed the *cumulative value* of such a life against time, that is, the value it would have if it ended at any given time, we would get a flat line. The slope of that graph, the extra cumulative value divided by the time already lived, would always be zero.[6]

To take a specific example, suppose that Zeno is added to some antecedent outcome A with a constantly neutral life $\boldsymbol{n}$ of 80 years. And then his lifespan is repeatedly halved, as in the following table, where the dash indicates Zeno's nonexistence.

[6] Compare Broome (2004: 68) and Brown (ms). In the end, on page 254, Broome defines "constantly neutral" in a way which presupposes coherence of temporal welfare, which my definition does not. And on pages 23-4 he also assumes discrete rather than continuous time. Brown's (ms) definition and framework are much closer to mine. His term for "constantly neutral" is "flatline".

	Zeno	People in A
A_1	Life $\mathscr{n}$	Unaffected
A_2	Life $\mathscr{n}$ cut short at 40 years	Unaffected
A_3	Life $\mathscr{n}$ cut short at 20 years	Unaffected
...	...	...
A	—	Unaffected

Table 5-1

Zeno's life $\mathscr{n}$ is just as good when it is lived in full as it is when cut short at 40 years, and just as good then as it is when cut short at 20 years, and so on. Hence, each outcome is just as good for Zeno as the next. Since others are unaffected, it is therefore plausible that each outcome is just as good as the next. This follows from

> **Pareto Indifference**. If the same number of people exist in outcomes X and Y, and everyone is equally well-off in X as in Y, then X and Y are equally good.

It is also plausible that the sequence of outcomes $A_1, A_2, ...$ converges to outcome A. That is, the sequence of outcomes where Zeno's life is getting shorter and shorter converges to an outcome where Zeno does not exist at all.

Why? First, the region of the spatiotemporal difference between A and members of the sequence is getting smaller and smaller. For i large enough, A_i differs from A only for a fraction of a second. Moreover, the difference does not "blow up" as it gets smaller and smaller spatiotemporally. The difference between A_i and A is the same as that between A_1 and A, except more localized.

To make the claim of convergence even more plausible we can assume that Zeno's life $\mathscr{n}$ is open on the left, so that there is no first moment of time when it is lived. It is therefore like the left-open interval, (0,1], containing all numbers between

zero and one, excluding zero, as opposed to a left-closed interval like [0,1], which also contains zero. Hence, if Zeno's life $\boldsymbol{n}$ were shortened to zero length, it arguably could not become a point-sized zero-length life, but would have to instead disappear altogether.

We can also think of $A_1, A_2, ...$ as the outcomes of successive finite stages of a *supertask*, a situation where infinitely many actions are performed in finite time, with A being the outcome at the ω-th stage, coming after all the finite stages. For example, compare Zeno's predicament with a version of Benardete's (1964: 259-60) "paradox of the gods" where a traveller wants to go from point 0 to point 1 but an infinity of impenetrable walls of decreasing thickness is put up in between, with the first wall halfway, at point ½, then halfway between 0 and ½, at point ¼, and so on.[7] The traveller therefore has to stay put at point 0. Similarly, if, say, we progressively tweaked Zeno's environment to make him more and more short-lived, his career would never get started.[8]

Given the claim that Zeno's life $\boldsymbol{n}$ is constantly neutral, Pareto indifference, and the claim of convergence, we therefore obtain the following pattern:

$$A_1 \sim A_2 \sim \cdots \to A.$$

---

[7] This version is discussed by Peijnenburg & Atkinson (2010). Unlike Benardete's original – discussed by Priest (1999) and Yablo (2000) – this one is unparadoxical, as Peijnenburg & Atkinson also show.

[8] I think both considerations of overall physical similarity and supertasks give us a handle on the notion of convergence. The latter might actually depend on the former. To determine what would happen at the ω-th stage of supertasks, authors such as Allis & Koetsier (1995) and Earman & Norton (1996) appeal to principles of physical continuity, thus presupposing physical convergence. Recent discussion of convergence and closeness in physics can be found in Fletcher (2020). If physics is too contingent for value theory, we might take inspiration from Lewis (1983) and define closeness and convergence in terms of perfectly natural properties instead.

Hence, by transfinite transitivity of equal goodness:

$$A_1 \sim A.$$

Since we assumed nothing special about Zeno or about other people in A, we can conclude, in general, that adding *any* constantly neutral and left-open life to *any* outcome is equally good as not adding it.[9]

We can immediately lift the restriction to left-open lives by using ordinary transitivity and Pareto indifference plus the extra assumption that *someone could* have a constantly neutral and left-open life.[10] That extra assumption is plausible enough. If time really is a continuum, why cannot someone's life occupy a temporal region open on the left?[11]

To see how the argument goes from there, consider the following table.

	Zeno	People in A
A	—	Unaffected
B	Life $\boldsymbol{n}$, left-open	Unaffected
C	Life $\boldsymbol{n}$, left-closed	Unaffected

Table 5-2

[9] Transfinite *acyclicity* of equal goodness gives a correspondingly weaker claim that the outcome of adding a constantly neutral life is neither better nor worse than the outcome of not adding it.

[10] This makes my argument unlike similar arguments in Bader (ms). Other differences include that Bader does not use constantly neutral lives, nowhere appeals to transfinite transitivity/acyclicity, and mostly works directly with permissible choice rather than value. Nonetheless, this paper would not have existed without the inspiration of Bader's earlier work.

[11] While plausible enough, it is part of an ancient debate about the boundaries of ordinary objects in continuous space; see Varzi (2015). Suffice to say, I have Leonardo da Vinci and Bertrand Bolzano on my side.

Think of outcome A as status quo. Zeno can then be added with a constantly neutral life $\boldsymbol{n}$ which will be left-open in outcome B and left-closed in outcome C.

First, if the argument above goes through for left-open lives, then A and B are equally good. The second step begins by noting that B and C are also equally good for Zeno. His life is longer by an instant in the latter outcome, but that should not matter since it is constantly neutral either way. And since other people are unaffected, it follows from Pareto indifference that B and C are equally good.

A stronger claim seems true as well: if someone's life in two outcomes is identical except that it is left-open in one and left-closed in the other, then both outcomes are equally good for that person. By analogy, recall that, in calculus, the area under a curve between two bounds a and b is the same whether or not the bounds themselves are included: a strip of no width can add nothing to total area.

Hence, A and B are equally good, and B and C are equally good. So, by transitivity of equal goodness, A and C are equally good. It does not matter whether a constantly neutral life is left-open or not. In general, we get

> **The Constant Addition Principle**. If outcomes X and Y differ only in that there is one extra person in Y living a constantly neutral life, then Y is equally good as X.

We can run direct arguments for similar principles on the negative and the positive side. For the former, consider adding a life which is *constantly bad* in the sense that it is worse the longer it is. If we graphed its cumulative value against time, we would get a downward-sloping curve. We can then imagine adding Zeno with such a life to outcome A and then repeatedly halving his lifespan. That each outcome in the resulting sequence is better than the last follows from a strengthening of Pareto indifference,

Strong Pareto. If the same number of people exists in outcomes X and Y, and everyone is at least as well-off in X as in Y, then X is at least as good as Y, and if, in addition, some are better-off in X than in Y, then X is better than Y.

Then the relevant claim of convergence and transfinite transitivity of strict worseness implies that adding Zeno with a constantly bad life is worse than not adding him.

On the positive side, the argument is similar. We make Zeno's life *constantly good*, so that it is better the longer it is. It then follows, this time using transfinite transitivity of strict betterness, that adding Zeno with such a life is better than not adding him. In either case the implicit restriction to left-open lives is easily removed as before.

We therefore obtain *negative* and *positive* versions of the constant addition principle, with "constantly neutral" and "equally good as" replaced by "constantly good" and "better than", and "constantly bad" and "worse than", respectively.[12]

3.2 Mere additions

These principles tell us something about adding lives with a uniform trajectory through time, *constant additions*, as it were. What do they imply about other *mere additions*?

[12] This matters as transitivity of indifference (the preferential counterpart of equal goodness) was historically the most readily rejected form of transitivity, for example, by Armstrong (1948: 3): "That indifference is not transitive is indisputable, and a world in which it were transitive is indeed unthinkable". See also Fishburn (1970).

We can use the special status of constantly neutral lives to answer this question. Say we are wondering whether it is better to add Zeno with some life ℓ to outcome A, as in the following table.

	Zeno	People in A
A	—	Unaffected
B	Life $\boldsymbol{n}$	Unaffected
C	Life ℓ	Unaffected

Table 5-3

That is, we are wondering whether C is better than A. To find out, it is enough compare life ℓ with life $\boldsymbol{n}$, a constantly neutral one. If ℓ is better than $\boldsymbol{n}$, then it follows, by strong Pareto, that C is better than B. Since, by the constant addition principle, B is equally good as A, it follows, by ordinary transitivity, that C is better than A. So, in general, it is always better to add a life that is better than some constantly neutral life. We can put this by saying that lives better than some constantly neutral life have *positive contributive value*. Analogously we can also show that lives worse than (just as good as) some constantly neutral life have *negative* (*zero*) *contributive value*.

In fact, a comparison with a constantly neutral life is not only sufficient but also necessary to determine a life's contributive value. Suppose, for example, that it is better to add Zeno with ℓ to A. That is, C is better than A. Since, by the constant addition principle, A is equally good as B, it follows by ordinary transitivity that C is better than B. But note that these two differ only in that Zeno has life ℓ in the former and life $\boldsymbol{n}$ in the latter. Since no one else is affected, it is plausible to

think that ℓ must therefore be better than $\boldsymbol{n}$.[13] It would then follow that if a life has positive contributive value, it must be better than a constantly neutral life. Analogously we can show that lives with negative (zero) contributive value are worse than (just as good as) some constantly neutral life.[14]

If we accept both directions of this argument, we see that constantly neutral lives neatly divide all possible lives in terms of their contributive value. In this way we know the status of all mere additions, not just constant additions.[15]

4 Implications for population ethics

Note that, so far, we did not have to assume that welfare can be numerically measured in any fine-grained manner, and we assumed nothing about what it takes for a life to be worth living or worth not living.

By contrast, theories and principles in population ethics typically assume something about both. First, they typically assume that welfare is interpersonally measurable on a ratio scale where, second, zero corresponds to the welfare level

---

[13] There are two ways this can be argued. First, if $\boldsymbol{n}$ is at least as good as ℓ, then it follows, by strong Pareto, that B is at least as good as C, contradicting the claim that C is better than B. Hence, ℓ is either better than $\boldsymbol{n}$ or incomparable with it. We can rule out the latter possibility by assuming that all lives are comparable in personal value. Second, we can use a same-number independence principle (as in Blackorby et al. (2005: 159)) without assuming comparability.

[14] This claim is compatible with some lives being incomparable with constantly neutral ones, as suggested by Gustafsson (2020). These lives would then have *undefined contributive value*: adding them would be incomparable with not adding them.

[15] The arguments of this section and the last also imply that all constantly neutral lives are equally good, worse than constantly good lives, and better than constantly bad lives.

of a life which is *neutral* in terms of personal value: on the boundary between worth living and worth not living.

In this framework, the key question becomes "What is the welfare level of constantly neutral lives?" Given the argument so far, it is tempting to accept what Brown (ms) would call

> **The Flatline Analysis of Neutrality**. A life is neutral iff it is just as good as a constantly neutral life.

Brown (ms) is sympathetic to this analysis and does much to defend it. Broome (2004: 68) suggests a similar analysis, too. Accepting it allows us to relate our conclusions so far with well-known theses in population ethics.

First, we get the nicely harmonious principle:

> **The Equivalence of Personal and Contributive Value**. The sign of a life's contributive value is equal to the sign of its welfare.[16]

It rules out theories for which the two signs can differ.[17] One example is

> **Average Utilitarianism**. Outcome X is at least as good as outcome Y iff average welfare in X is at least as high as average welfare in Y.

This is because if the average welfare of an antecedent population is negative enough, adding an extra life with negative welfare might increase average.[18] Another example is

---

[16] Compare Gustafsson (2020: 7), who traces this sort of principle back to Rabinowicz (2009).

[17] If we used transfinite *acyclicity* principles instead of transfinite *transitivity* principles, we would get the weaker claim that the sign of a life's contributive value is *not distinct* from the sign of its welfare, which allows for, say, a positive-welfare life with undefined contributive value.

[18] Compare Parfit's (1987 [1984]: 422) Hell Three.

> **Critical-Level Utilitarianism**. Outcome X is at least as good as outcome Y iff total welfare in X is at least as high as in Y after positive constant α, the critical-level parameter, was subtracted from each person's welfare in both X and Y.[19]

Hence, a life might have positive welfare but contribute negatively to the value of the outcome if it happens to be below α.[20]

Another upshot is that we can quickly get to

> **The Repugnant Conclusion**. For any outcome in which many people exist, at a high welfare level, there is a better outcome in which many more people exist, at a barely positive welfare level.[21]

To see this, start with some outcome A where many people exist at a high welfare level. Keep adding constantly neutral lives until we get an outcome, say, B, where average welfare is barely positive. By the constant addition principle, each such addition preserves overall value. Hence, by ordinary transitivity, all of them do. So, B is equally good as A. Let Z be an outcome which is like B except that total welfare is slightly higher and equally distributed but where everyone's welfare is still barely positive. That Z is better than B follows from

> **Non-Anti-Egalitarianism**. If outcome X has higher total welfare, higher average welfare, and is more equal than outcome Y, then X is better than Y.[22]

---

[19] Compare Broome (2004: 255) and Blackorby, Bossert & Donaldson (2005: 137-8).

[20] This leads to the Sadistic Conclusion of Arrhenius (2000: 251).

[21] Compare Parfit (1987 [1984]: 388)

[22] See Ng (1989: 238).

Hence, Z is better than B, B is equally good as A, so, by transitivity, Z is better than B. This is the repugnant conclusion.[23]

In fact, once welfare is interpersonally measurable on a ratio scale, we can draw some important conclusions even without assuming the flatline analysis of neutrality.

The argument for the repugnant conclusion, for example, still works if "barely positive welfare level" is everywhere replaced by "welfare level barely above that of constantly neutral lives". If the original repugnant conclusion is repugnant, it should also be repugnant with that replacement: the Z outcome can differ little from one where everyone's life is on the verge of being always worth ending.

Lastly, given the argument about constant additions, we can reduce comparisons between outcomes with different population sizes to comparisons between outcomes with the same population size. For example, consider the following table, showing people's welfare levels, with $w(\boldsymbol{n})$ being the welfare level of a constantly neutral life $\boldsymbol{n}$ (not necessarily zero, unless we assume the flatline analysis).

[23] This argument is similar to Parfit's (1987 [1984]: 419-441) mere addition paradox. We can see that claims about convergence, transfinite transitivity principles and Pareto principles can together replace Parfit's mere addition principle, given the flatline analysis of neutrality. This is how the current paper's argument can be read as an impossibility theorem, contributing to the literature exemplified by Ng (1989), Arrhenius (2000), and Blackorby et al. (2005: 180-208).

	Zeno	Zelda	Xenon
A	10	—	—
B	5	5	5
C	10	$w(\boldsymbol{n})$	$w(\boldsymbol{n})$

Table 5-4

How do A and B compare? From the constant addition principle and Pareto indifference, it follows that A and C are equally good. Hence, by transitivity, A and B compare in the same way that C and B compare. For example, according to

> **Same-Number Utilitarianism**. Outcome X is at least as good as outcome Y iff total welfare in X is at least as high as total welfare in Y, provided that X and Y have the same population size.

Then B is better than C iff

$$3 \times 5 > 10 + 2 \times w(\boldsymbol{n}),$$

which, subtracting $3v(\boldsymbol{n})$ from both sides, becomes

$$3 \times \big(5 - w(\boldsymbol{n})\big) > 10 - w(\boldsymbol{n}).$$

So, in general, outcomes compare according to totals of welfare less the welfare level of a constantly neutral life. This is a version of critical-level utilitarianism. Given the flatline analysis of neutrality, $w(\boldsymbol{n})$ becomes zero and we get *total utilitarianism* which simply removes the same-size proviso in same-number utilitarianism. A similar story is true for other same-number principles, such as egalitarianism, prioritarianism, maximin, leximin, and so on.

5 Objections and loose ends

I will now consider a number of objections and loose ends that arise regarding the basic argument of the last section, about additions of constantly neutral lives open on the left. First: doesn't the argument overgenerate by implying the contentious claim that a life can be compared with nonexistence in terms of personal value? Second: is it even possible to make a life arbitrarily short in length? And third: can we really make sense of the idea that a sequence of outcomes with some number of people converges to an outcome with a different number of people?

5.1 Comparativism

I argued that adding Zeno with a constantly neutral life to *any* outcome A is just as good as not adding him. What if A is otherwise empty, as in the following table?

	Zeno	People in A
A_1	Life $\boldsymbol{n}$	—
A_2	Life $\boldsymbol{n}$ cut short at 40 years	—
A_3	Life $\boldsymbol{n}$ cut short at 20 years	—
…	…	…
A	—	—

Table 5-5

The argument about constant additions still leads to the conclusion that A_1 is equally good as A. But this does not show that outcome A_1 is equally good *for Zeno* as outcome A. Hence, the argument establishes

General-Value Comparativism. Outcomes where one does not exist can be comparable in terms of general value with outcomes where one does exist.

But not necessarily

Personal-Value Comparativism. Outcomes where one does not exist can be comparable in terms of personal value with outcomes where one does exist.[24]

To show the latter we would need to use

Transfinite Transitivity of Equal Goodness-For. For all value-bearers $X, X_1, X_2, ...$, the sequence $X_1, X_2, ...$ converges to X, and X_1 is equally good for one as X_2, X_2 is equally good for one as X_3, ..., then X_1 is equally good for one as X.

Recall that value-bearers are things that are comparable with something in terms of the relevant value relation, in this case the "equally good-for" relation. And unless we already accept personal-value comparativism, we will not think that A, an outcome where Zeno does not exist, is a value-bearer in that sense.

The restriction to value-bearers is redundant in the case of ordinary (finite) transitivity where the antecedent already implies that all of the items at issue are value-bearers in the relevant sense. This is not true in the case of *transfinite* transitivity, since the limit outcome is not mentioned in the antecedent at all. There we need to add that restriction explicitly.[25]

---

[24] See Broome (1993) for a classic argument against personal-value comparativism, and Bykvist (2007) for a more recent discussion.

[25] Even though we lack a non-question-begging argument for personal comparativism, we have a non-question-begging argument for the conditional: if personal-value comparativism, then

5.2 Divisibility

In the last section I also assumed that a life can be made arbitrarily short in length. This is to assume something both about the nature of time and the nature of life: that there is no smallest interval of time, and that there is no smallest possible lifespan.

To see that the former is actually needed, suppose that there is a smallest interval of time, say, Δ units. Then we cannot make Zeno's lifespan arbitrarily close to zero since it cannot go down below Δ units. Hence, no sequence of outcomes where Zeno's life gets shorter and shorter converges to an outcome where Zeno does not exist at all.

In response, we can say that continuous time is routinely assumed in physics and that ethics should be no worse-off in that respect.[26] So, while the possibility of discrete time is a real challenge, it is not a pressing one.

On the other hand, the possibility that there is a limit on how short a life can be, biologically or psychologically speaking, is less speculative. Yet it was also implicitly assumed away in last section's argument. Luckily, it can be accommodated, at least to an extent.

---

nonexistence is just as good for one as existence with a constantly neutral life. Compare Nebel (2019: 325).

[26] Compare, for example, Pruss (2018: 172): "Standard formulations of major physics theories from Newton onwards either model time with the real numbers or model spacetime as a continuous manifold with local coordinates (…) The continuity involved is essential to the differential equations in which the laws of physics are couched."

To see this, note that there are two ways to read the claim that Δ units, say, is a minimum duration of life:

(i) Any possible life lasts at least Δ units of time.

(ii) Any possible life lasts more than Δ units of time.

The difference is subtle but important. According to the first, the temporal interval $[0, \Delta]$ is long enough to fit a life: if something starts at time 0 and lasts until time Δ inclusive, it lasts for Δ units in total. But, according to the second, that is too short. To be a life, something needs to last longer than any non-life *by a positive margin.*

The last section's argument carries over with small changes on the second reading but fails on the first. To see this, consider the following table.

	Zeno	People in A
A_1	Life $\boldsymbol{n}$, 30000 days long	Unaffected
A_2	Life $\boldsymbol{n}$ cut short at 15000.5 days	Unaffected
A_3	Life $\boldsymbol{n}$ cut short at 7500.75 days	Unaffected
A_4	Life $\boldsymbol{n}$ cut short at 3750.875 days	Unaffected
...	...	...
A	Life $\boldsymbol{n}$ cut short at 1 day	Unaffected

Table 5-6

Let's assume, for example, that Δ units is 1 day. A_1 is the outcome of adding Zeno to A with a constantly neutral life $\boldsymbol{n}$ of 30,000 days, which is about 82 years. Then if in some subsequent outcome Zeno's lifespan is T days, we cut it down to $1 + \frac{T-1}{2}$ days in the next. Arguably, the sequence of outcomes $A_1, A_2, ...$ converges to A, where Zeno's life is cut short at exactly the 1 day mark.

On the first reading, Zeno has a life in all outcomes in the sequence as well as in the limit outcome. Hence, using last section's strategy, it follows, at best, that

adding a constantly neutral life is just as good as adding a constantly neutral life of minimal duration. But, on the second reading, Zeno does not exist in A. Put differently, "life $\boldsymbol{n}$ cut short at 1 day" does not pick out a life but perhaps something which could have been a life if it lasted a little longer. Hence, we can argue as before that adding a constantly neutral life is just as good as not adding it at all.

So, to mount a successful challenge to last section's argument, one would have to argue not only that there is a minimum duration of life, but also for a specific reading of "minimum duration". But the two reading seem on a par, with the second reading having a slight intuitive edge. So, we again have a real challenge but not a pressing one.

5.3 Distance metric

But can we even make sense of convergence across population sizes? In some cases this is intuitive enough. But to improve our grip on it, I will now give a toy model of closeness and, so, convergence. It is meant as a consistency check, and perhaps a starting point for more serious models.

In the extant population ethics literature, convergence is sometimes used in the sense of convergence in welfare profile.[27] More precisely, it is often said that the sequence of outcomes $X_1, X_2, ...$ converges to outcome X if the welfare profiles of $X_1, X_2, ...$ converge to that of X, where a *welfare profile* is an n-tuple of real numbers representing the welfare of the n people existing in a given outcome. This assumes, of course, that welfare can be measured on a real-valued scale.

[27] See Blackorby et al. (2001) and Broome (2003).

Since a welfare profile is a point in a Euclidean n-space, $\mathbb{R}^n$, convergence between welfare profiles can be fixed by the *Euclidean distance*, so that the distance between welfare profiles $(w_1, w_2, \ldots, w_n)$ and $(u_1, u_2, \ldots, u_n)$ is given by:

$$\sqrt{(w_1 - u_1)^2 + (w_2 - u_2)^2 + \cdots + (w_n - u_n)^2},$$

For example, on the real plane ($\mathbb{R}^2$), the Euclidean distance between two points is simply the length of a line segment joining them.

But if two outcomes differ in population size, their welfare profiles are of different dimensions. Hence, their Euclidean distance is undefined, unless we assume the controversial personal-value comparativism.[28] But, for our purposes, there is an easy fix: include information about lifespans in addition to welfare levels.

To see how it goes, think of a life as a point in two-dimensional space, the x-axis giving its length, the y-axis giving its overall welfare. The space will look as follows.

[28] It would also be unwise to *assume* it here, since it could make last section's arguments redundant. See, for example, Holtug (2001: 363-364).

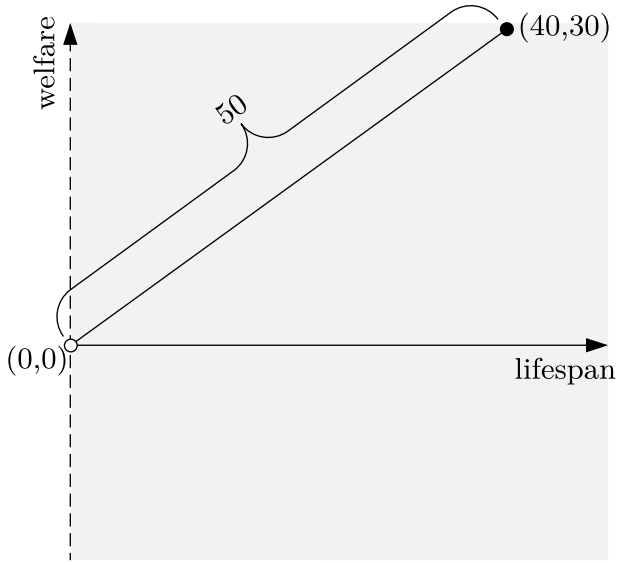


Figure 5-1. Distance metric

To prop up last section's argument, we need to make sense of the idea that, as someone's lifespan goes to zero, the whole outcome gets closer to one where they do not exist. But, recall, we only need to make sense of this for lives which are left-open. Hence, we can ignore the possibility that someone's life could correspond to a point directly on the y-axis (which is therefore dotted).

So, intuitively, if we make someone's life closer to the y-axis, we will move the whole outcome closer to one without them. But to get a well-defined distance, we need to pick a single point on the y-axis as reference. Here it is natural to pick the origin, (0,0).

So, for each person, we can now measure the distance to an outcome without them by simply using the Euclidean distance in a welfare/lifespan space like in the figure above. For example, the distance between (0,0) (representing nonexistence) and (40,30) (a life of 40 years with welfare 30) is easily calculated to be 50. We can now aggregate these individual measures into a single measure for whole outcomes.

More precisely, we can define a *welfare-lifespan profile* to be a (countably infinite) list of pairs of welfare levels and *lifespans* for each *possible person*, including possibly the (0,0) pair (representing nonexistence). Examples are $((w_1, l_1), (w_2, l_2), \dots)$ and $((u_1, k_1), (u_2, k_2), \dots)$, with $w_1, w_2, \dots, u_1, u_2, \dots$ representing welfare levels and $l_1, l_2, \dots, k_1, k_2, \dots$ representing lifespans. Then we can define the distance between their corresponding outcomes to be:

$$\sqrt{\begin{array}{c}(w_1 - u_1)^2 + (l_1 - k_1)^2 \\ +(w_2 - u_2)^2 + (l_2 - k_2)^2 \\ +\cdots\end{array}}$$

This new formula can easily be verified to be a distance metric. In cases where population size is finite, it is also always well-defined, since at most a finite number of its terms will be nonzero. In fixed-population fixed-lifespan comparisons, it reduces to the Euclidean distance between welfare profiles. And, lastly, it implies the convergence claims needed for last section's arguments.

This toy model also has no untoward substantive implications. For example, even though one's closeness to nonexistence is measured from the origin, (0,0), it does not follow that nonexistence is a bearer of personal value in any sense, and a zero-welfare life is not automatically treated as nonexistence (since it has positive lifespan).

6 Procreative asymmetry

All arguments in this paper so far were about evaluating outcomes where different numbers of people exist. I will now consider the issue of choosing between such outcomes. In this context, transfinite *acyclicity* becomes more important than transfinite *transitivity*. In particular, transfinite acyclicity of *strict worseness* rather than *strict betterness* turns out to be crucial. This asymmetry between worseness and betterness has significant implications for population ethics and can lead to a form of *weak procreative asymmetry*, according to which, other things equal, adding bad lives is more objectionable than not adding good ones.[29]

6.1 Asymmetry of transfinite cycles

Since transfinite transitivity/acyclicity are infinitary principles, to see how they bear on permissible choice, we have to consider choices from infinite sets of outcomes. But infinite sets of outcomes are notoriously problematic.[30] Suppose, for example, the more money the better and consider $\{\$1, \$2, \$3, ...\}$, the set of natural-valued dollar amounts, or $[\$0, \$100)$, the set of dollar amounts greater than \$0 but strictly less than \$100. In either case no outcome is *maximal* in the sense of not being worse than any available outcome. *A fortiori*, no outcome is *best* in the sense of being at least as good as any available. But it is often assumed

[29] The procreative asymmetry first appears in Narveson (1973), the label itself being due to McMahan (1981). In its stronger form it says that, other things equal, adding bad lives is objectionable and not adding good lives is *not at all* objectionable. Weak procreative asymmetry is defended in McMahan (2009).

[30] See Pollock (1983), Slote (1989: 47-81), Sorensen (1994), and Arntzenius et al. (2004).

that, setting deontological considerations aside, one is permitted to choose all and only maximal outcomes. So, it is puzzling what one is to choose in these cases.

Here the puzzlement arguably arises because the set of available outcomes, S, fails to have the property of

> **Compactness**. Every sequence drawn from S has a subsequence which converges to something in S.

In our examples, the sequence $\$1, \$2, \$3, ...$ diverges to infinity, so has no convergent subsequence, and even though the sequence \$0, \$50, \$75, \$92.5, … converges to \$100, \$100 itself is not an available outcome. Since many of the paradigmatically problematic infinite decision puzzles have to do with noncompactness, it is reasonable to expect compact sets of outcomes to be unproblematic when it comes to the existence of maximal (generally: choiceworthy) outcomes.

Hence, it is important news that transfinite cycles of *strict worseness* rule out the existence of maximal elements in compact sets. By contrast, transfinite cycles of *strict betterness* do not need to do so.

For example, the former sort of cycle can arise if adding constantly bad lives is sometimes better than not adding them (a potential implication of average utilitarianism given the flatline analysis of neutrality). To see this, suppose we add Zeno with a constantly bad life $\mathscr{b}$ to some antecedent outcome A and then progressively halve his lifespan.

	Zeno	People in A
A_1	Life $\mathscr{b}$	Unaffected
A_2	Life $\mathscr{b}$ cut short at 40 years	Unaffected
A_3	Life $\mathscr{b}$ cut short at 20 years	Unaffected
…	…	…
A	—	Unaffected

Table 5-7

If adding Zeno with $\mathscr{b}$ is better than not adding him, then, with strong Pareto, we get the following transfinite cycle of strict worseness:

$$A \prec A_1 \prec A_2 \prec \cdots \rightarrow A.$$

Let $S = \{A, A_1, A_2, \ldots\}$. No outcome is maximal in S, as no outcome of the form A_i can be maximal, since it is worse than A_{i+1}, and A cannot be maximal either, since it is worse than A_1. Yet S is easily verified to be compact.

A mirror image of this situation can arise if adding constantly good lives is sometimes worse than not adding them (a potential implication of critical-level utilitarianism given the flatline analysis of neutrality). In the following table, $\mathscr{g}$ is some constantly good life.

	Zeno	People in A
B_1	Life $\mathscr{g}$	Unaffected
B_2	Life $\mathscr{g}$ cut short at 40 years	Unaffected
B_3	Life $\mathscr{g}$ cut short at 20 years	Unaffected
…	…	…
B	—	Unaffected

Table 5-8

If adding Zeno with $\not g$ is worse than not adding him, then, with strong Pareto, we get the following transfinite cycle of strict betterness:

$$B \succ B_1 \succ B_2 \succ \cdots \to B.$$

Let $S = \{B, B_1, B_2, \dots\}$. If we assume ordinary (finite) transitivity, S does have a maximal, even best, outcome, namely, B. Hence, we see that the two sorts of transfinite cycles (and the theories that can generate them) have different implications for the possibility of maximizing choice, even from compact sets.

6.2 Asymmetry of pumpability

This difference maps on to whether the two sorts of transfinite cycles make one liable to money pumps. We say that an agent is liable to *forcing money pumps* if in some possible situation they *have to* make a series of actions so that they end up paying some cost even though an otherwise identical costless alternative could also be obtained, and that they are liable to *nonforcing money pumps* if in some possible situation they *might* but perhaps do not *have to* make such a series of actions.[31]

We can leave the nature of this cost open. It need not be monetary but can instead denominated in some other relevant currency. And it need not fall on the agent alone but can instead be spread to some or all of the people involved. The basic idea behind money-pump arguments is that, whatever the cost is, one's moral theory would not tell one to pay it in advance. Hence, if, together with the

[31] The distinction appears in Gustafsson & Espinoza (2010).

structure of one's decision problem, following one's theory means that one ends up paying anyway, then the theory is self-defeating in an objectionable manner.[32]

Let's start with finite cycles. Suppose that A is worse than B, which is worse than C, which is worse than A, and consider the following decision tree where squares represent the agent's choices, leading to further choices or to final outcomes.[33]

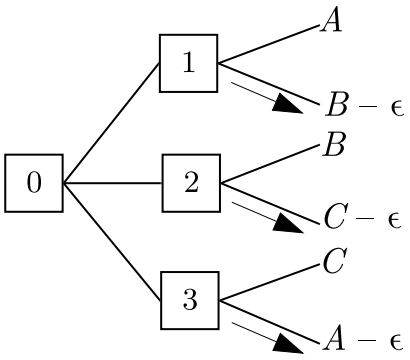


Figure 5-2. Cantwell pump

Here we offer the agent a choice from A, B, C, and, whatever they choose, we give them a chance to switch to the next-better outcome for a small payment. The claim is that the agent, if guided by value, will end up paying no matter what (as indicated by the arrows). Why?

First, the agent has to choose B over A. This follows if the agent is guided by value in a minimal way, namely, in pairwise choices. But then, since B is strictly better than A, the agent should choose B even at some small cost ϵ. Similarly for

[32] This is not the place for a full defence of money-pump arguments in ethics. For a recent sympathetic discussion, see Gustafsson (2015).

[33] This decision tree is due to Cantwell (2003: 389) and also employed by Gustafsson (2015: 1596-1597). Money pumps first appear in Davidson et al. (1955).

C and B, and A and C. Hence, they have to go down at all nodes following the initial one.

Second, the agent is permitted to do something at the initial node. From that node's vantage point there are only finitely many outcomes that can be obtained. And, as is often assumed, finite sets of outcomes never present moral dilemmas.[34] This is plausible so long as deontological considerations are set aside. So, the agent is permitted to embark on a plan to get one of the obtainable outcomes.

The agent is therefore permitted to go somewhere from the initial node but then they have to go down. Hence, in two moves, the agent is required to pay for something they could have had for free. This is a forcing money pump. Since this is objectionably self-defeating, we can conclude that strict worseness cannot be cyclic after all.[35]

Now consider an infinite version of this pump, directed against the transfinite cycle of strict worseness from Table 5-7.

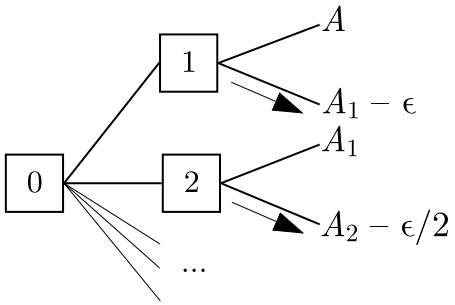


Figure 5-3. Infinite Cantwell pump 1

[34] Compare Kreps's (2013: 3) assumption of finite nonemptiness in rational choice theory.

[35] There are ways to get around this pump, for example, by adopting *resolute choice*. See Machina (1989) and McClennen (1990). I lack space to defend Cantwell's pump against this possibility.

Here we offer the agent a choice between outcomes $A, A_1, A_2, ...$, and, whatever they choose, we give them a chance to switch to the next-better outcome for a small payment.[36] The claim is, again, that the agent (an average utilitarian, say), if guided by value, will end up paying no matter what.

The first step is like before (this time backed up by strong Pareto). The second step appeals not to the claim that finite sets never present moral dilemmas, but instead to the claim that *compact* sets do not. While choice form infinite sets of outcomes might be hard, we saw above that many paradigmatic infinite decision puzzles are due to noncompactness. Hence, if we demand choice to be possible from finite sets, we should be happy to demand it from infinite compact sets.[37] In the current example, it also does not *seem* like the agent's initial choice is a moral dilemma: they are effectively asked whether to add Zeno with a constantly bad life and, if so, how long to make it.

Hence, in two moves, the agent is again required to pay for something they could have had for free. This is a forcing money pump. If we should reject cycles of strict worseness in response to Cantwell's original pump, we should reject them here, too.

By contrast, the transfinite cycle of strict betterness from Table 5-8 does not lead to trouble in the same sort of pump.

---

[36] The side payments converge to zero to ensure compactness of the set of obtainable outcomes. It is not necessary that they decrease *geometrically*.

[37] Compactness is routinely assumed in economics in this context, see Kreps (2013: 1-29). And, in mathematics, compactness is seen as a generalization of finiteness, as per Hermann Weyl's apocryphal gloss on one of its many equivalent formulations: "if a city is compact, it can be guarded by a finite number of arbitrarily near-sighted policemen", see Hewitt (1960). I say more in chapter 6.

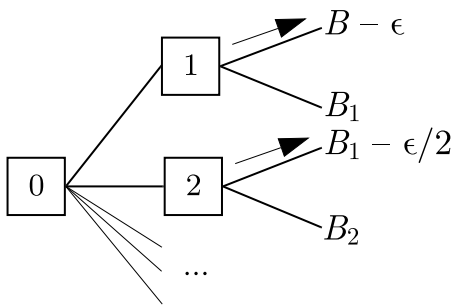


Figure 5-4. Infinite Cantwell pump 2

In this case, the agent should arguably end up paying ϵ for B, since B is best in $\{B, B_1, B_2, ...\}$. Since B is not available for free anyway, the agent does not end up paying for something they could have had for free. And if B were also freely available, say, directly from the initial node, the agent should arguably go for that free B.[38]

6.3 Asymmetry of procreation

What makes one liable to money pumps in this context is not the transfinite cycle of strict worseness itself but rather the transfinitely cyclic pattern of pairwise choices it generates. And that pattern can arise independently if, for example, adding bad lives is sometimes permissible. By contrast, it does not immediately arise if not adding good lives is sometimes permissible.

To get the argument going, we need one extra plausible assumption, namely, that if adding a *bad* life is permissible, then adding an equally good but *constantly bad* life is just as permissible. Without loss of generality, then, suppose that it is

[38] This does not, strictly speaking, show that *no* pump is available against a transfinite cycle of strict betterness, although I think that stronger claim is true as well. See my "Pumping discontinuity" (unpublished).

permissible to add Zeno with constantly bad life 𝒷 to some antecedent outcome A, as in Table 5-7.

Now consider the corresponding decision tree in Figure 5-3. The claim is that if the agent is allowed to add Zeno with 𝒷 to A, then they *might* end up paying no matter what. (If they are *required* to add him, then they *have to* pay.)

The first step is that the agent may go down at all nodes following the initial node. First, the agent should go down at nodes 2, 3, 4, … because that is at least as good for all and better for some. This can be backed up by strong Pareto, this time deontically construed. Put differently, it is worth paying something to make Zeno's miserable life twice as short as it would otherwise be. And, second, it arguably follows from our hypothesis that it is permissible to add Zeno with 𝒷 to A that the agent *may* go down at node 1 as well.[39]

The next step is the same as before: either compact sets present no moral dilemmas, or this particular choice is not a moral dilemma.

Hence, the agent is permitted (at worst, required) to end up paying for something they could have had for free. This is a nonforcing money pump.

By contrast, the permission not to add a good life does not land our agent in trouble in the same way. This can be seen in the decision tree in Figure 5-4, on the supposition that the agent is permitted not to add Zeno with constantly good

[39] If, on balance, one's reasons favour adding Zeno, then that will be true if adding Zeno comes at a small cost. If, on balance, one's reasons favour adding and not adding equally, then we can arrange things so that one's reasons favour adding Zeno more (by making his life slightly better, say). And if the balance of reasons fails to decide, then, plausibly, making Zeno's creation slightly costlier will not change that.

life $\mathscr{g}$. There the agent is arguably still be permitted to end up with $B - \epsilon$, hence avoiding a money pump.

If arguments in the first part of this paper, based on transfinite transitivity, are sound, then adding a bad life and not adding a good life are both suboptimal: they make things worse. What we can conclude from this section, however, is that there is a special reason against adding bad lives which does not speak against not adding good lives. That is, the permission to do the former can make one liable to money pumps in a way in which the permission to do the latter does not. Hence, we get a form of the weak procreative asymmetry: adding bad lives is more objectionable than not adding good ones, other things equal. Importantly, this asymmetry at the level of choice can be combined with a fully symmetric picture at the level of value.[40]

7 Continuity

Transfinite transitivity and acyclicity principles are *topological principles* in the sense that they make an essential reference to the *topology* of the domain of evaluation such as facts about convergence and closeness. To forestall confusion, it is important to distinguish them from some better-known topological principles,

[40] This account of the asymmetry relies on an intrinsic asymmetry between infinite chains of worseness and betterness. It is therefore not *ad hoc*, unlike many accounts surveyed in Roberts (2011). It also does not rely on substantive deontological resources. It is close in spirit, and inspired by, the account in Bader's (ms), although Bader does not appeal to transfinite cycles nor money pumps but instead to principles of universalizability.

namely, continuity principles. This comparison also serves to underscore the former's appeal.[41]

I will focus on

> **Transfinite Transitivity of Weak Worseness**. For all value-bearers $X, X_1, X_2, ...$, if the sequence $X_1, X_2, ...$ converges to X, and X_1 is at least as bad as X_2, X_2 is at least as bad as X_3, …, then X_1 is at least as bad as X.

Put differently: for all value-bearers $X, X_1, X_2, ...$,

$$X_1, X_2, ... \to X \text{ and } X_i \preccurlyeq X_{i+1}, \text{ for all } i, \text{ implies } X_1 \preccurlyeq X.$$

Contrast this with

> **Continuity of Weak Worseness**. For all value-bearers $Y, X, X_1, X_2, ...$, if the sequence $X_1, X_2, ...$ converges to X, and Y is at least as bad as X_1, Y is at least as bad as X_2, …, then Y is at least as bad as X.

Put differently: for all value-bearers $Y, X, X_1, X_2, ...$,

$$X_1, X_2, ... \to X \text{ and } Y \preccurlyeq X_i, \text{ for all } i, \text{ imply } Y \preccurlyeq X.$$[42]

The idea is that value comparisons (here: weak worseness) are preserved in the limit.

Continuity and transfinite transitivity differ in two ways. First, the $X_1, X_2, ...$ sequence has to be $\preccurlyeq$-ordered for transfinite transitivity but not for continuity. Second, in the case of transfinite transitivity but not continuity, the Y against

---

[41] Another principle of this sort is *hypersensitivity avoidance*, discussed in Pummer (2019).

[42] In metric spaces, this is the same as the set $\{Y: Y \preccurlyeq Z\}$ being closed for all value-bearers Y. See, for example, Kreps (2013: 13-14) in microeconomics, and Blackorby et al. (2005: 92) in population ethics.

which members of the sequence are compared has to be the sequence's first member. So, it is easy to see that, given ordinary transitivity, continuity implies transfinite transitivity.[43] Yet the converse fails in an interesting way, since transfinite transitivity, unlike continuity, can accommodate lexical value relations.[44]

To see this, let's use Thomas's (2018: 813) toy example of *total lexical utilitarianism*. To introduce it, first assume that only two things matter for personal value: *love* and *money*. We will think of them as real-valued quantities. Then according to

> **Total Lexical Utilitarianism**. Outcome X is at least as good as outcome Y iff total love in X is greater than total love in Y, or they are equal and total money in X is at least as great as total money in Y.

So, each life and each outcome maps to a point of a two-dimensional space with total quantities of love on the x-axis, and of money on the y-axis. The horizontal dimension is lexically superior. Since we are in a two-dimensional space, let's assume that the relevant notion of convergence between outcomes is fixed by Euclidean distance.

---

[43] See Proposition 3.6 in Smith (1974) and my discussion in chapter 6. Bartha et al. (2014: 641-642, 657-658) also discuss the connection between principles they call "transfinite transitivity" and "continuity". But by "continuity" they mean continuity of the *numerical representation* of the underlying binary relation. As Luce & Suppes (1965: 265) indicate, continuity of the binary relation is the more fundamental notion.

[44] Compare Debreu (1953). Transfinite transitivity can also accommodate some incomplete value relations that continuity cannot. Compare Aumann (1962: 450–453).

It is easy to see that all this implies that Thomas's toy theory violates continuity of weak worseness, as in the following table.[45]

	Zeno (love, money)	Other people
A_1	(80, 0)	Unaffected
A_2	(40, 0)	Unaffected
A_3	(20, 0)	Unaffected
...	...	...
A	(0, 0)	Unaffected
B	(0, 100)	Unaffected

Table 5-9

In this example, Zeno's money-deprived life is diminishing in terms of love over the sequence $A_1, A_2, ...$ But so long as Zeno's life holds some love, each of these outcomes is better than B, where he has no love but some money. Yet the sequence $A_1, A_2, ...$ converges to A, where Zeno has no love and no money, which is therefore worse than B.

This is a violation of continuity of weak worseness but not transfinite transitivity of weak worseness. For the latter we would need not only the fact that B is worse than A_1 without being at least as bad as A, but also that all outcomes in the sequence $A_1, A_2, ...$ are at least as bad as the next, which we do not have in this case.

Thus, while proponents of lexical views in population ethics have reasons to be wary of topological principles like continuity, they have no immediate reason to

[45] Thomas (2018: fn. 18) says otherwise because he assumes the discrete topology rather than the more natural Euclidean one.

be wary of transfinite transitivity principles. Therefore, framing this paper in terms of transfinite transitivity has a real advantage, besides pumping transitivity-friendly intuitions, even though, logically speaking, much of it would work if we used continuity principles instead.[46]

8 Conclusion

This paper introduced and explained infinite extensions of transitivity and acyclicity in the context of population ethics. It showed how to use these principles (given a suitable notion of convergence or closeness) to establish that adding constantly neutral lives is just as good as not adding them, adding constantly good lives is better, and adding constantly bad lives is worse. These arguments have important implications for population ethics: ruling out a number of theories, feeding into arguments for the repugnant conclusion, and reducing different-number comparisons to same-number comparisons. It also addressed a number of challenges to these arguments: the issue of personal-value comparativism, the possibility of minimal quanta of time and life, and provided a toy model of convergence across population sizes. The paper also uncovered an asymmetry between infinite chains of betterness and worseness which leads to an asymmetry in liability to money pumps and to an asymmetry in the ethics of procreation. Transfinite transitivity and acyclicity were also argued to be more appealing than standard topological principles of continuity.

[46] This is especially true as transfinite transitivity of equal goodness happens to be equivalent to continuity of equal goodness. See Proposition 3.5 in Smith (1974).

9 References

Allis, V., & Koetsier, T. (1995). On some paradoxes of the infinite II. *The British Journal for the Philosophy of Science, 46*(2), 235-247. doi:10.1093/bjps/46.2.235

Armstrong, W. E. (1948). Uncertainty and the utility function. *The Economic Journal, 58*(229), 1-10. doi:10.2307/2226342

Arntzenius, F., Elga, A., & Hawthorne, J. (2004). Bayesianism, infinite decisions, and binding. *Mind, 113*(450), 251-283. doi:10.1093/mind/113.450.251

Arrhenius, G. (2000). An impossibility theorem for welfarist axiologies. *Economics and Philosophy, 16*(2), 247-266. doi:10.1017/S0266267100000249

Aumann, R. J. (1962). Utility theory without the completeness axiom. *Econometrica, 30*(3), 445-462. doi:10.2307/1909888

Bader, R. (ms). *The Asymmetry*. Unpublished manuscript.

Bartha, P., Barker, J., & Hájek, A. (2014). Satan, saint peter and saint petersburg. *Synthese, 191*(4), 629-660. doi:10.1007/s11229-013-0379-9

Benardete, J. A. (1964). *Infinity: An essay in metaphysics*. Oxford: Clarendon Press.

Birchenhall, C. R. (1977). Conditions for the existence of maximal elements in compact sets. *Journal of Economic Theory, 16*(1), 111-115. doi:10.1016/0022-0531(77)90126-0

Blackorby, C., Bossert, W., & Donaldson, D. (2001). Population ethics and the existence of value functions. *Journal of Public Economics, 82*(2), 301-308. doi:10.1016/S0047-2727(00)00135-3

Blackorby, C., Bossert, W., & Donaldson, D. J. (2005). *Population issues in social choice theory, welfare economics, and ethics*. Cambridge University Press.

Broome, J. (1993). Goodness is reducible to betterness: The evil of death is the value of life. In P. Koslowski, & Y. Shionoya (Eds.), *The good and the economical* (pp. 70-84). Springer Verlag.

Broome, J. (2003). Representing an ordering when the population varies. *Social Choice and Welfare, 20*(2), 243-246. doi:10.1007/s003550200175

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

Brown, C. (ms). *Better than nothing.* Unpublished manuscript. https://philpapers.org/rec/BROHTL

Bykvist, K. (2007). The benefits of coming into existence. *Philosophical Studies, 135*(3), 335-362. doi:10.1007/s11098-005-3982-x

Cantwell, J. (2003). On the foundations of pragmatic arguments. *Journal of Philosophy, 100*(8), 383-402. doi:10.5840/jphil2003100826

Carosi, L., & Zaffaroni, A. (1999). On the existence of maximal elements for partial preorders. *Journal of Information and Optimization Sciences, 20*(2), 271-286. doi:10.1080/02522667.1999.10699417

Davidson, D., Mckinsey, J. C. C., & Suppes, P. (1955). Outlines of a formal theory of value, I. *Philosophy of Science, 22*(2), 140-160. doi:10.1086/287412

Debreu, G. (1954). Representation of a preference ordering by a numerical function. In R. M. Thrall, C. H. Coombs & R. L. Davis (Eds.), *Decision processes* (pp. 159-166). New York: Wiley.

Earman, J., & Norton, J. (1996). Infinite pains: The trouble with supertasks. In Adam Morton, & Stephen P. Stich (Eds.), *Benacerraf and his critics* (pp. 231-261). Blackwell.

Fishburn, P. C. (1970). Intransitive indifference in preference theory: A survey. *Operations Research, 18*(2), 207-228. doi:10.1287/opre.18.2.207

Fletcher, S. C. (2020). The principle of stability. *Philosophers' Imprint,* 20(3), 1-22.

Gillies, D. B. (1959). Solutions to general zero-sum games. In A. Tucker, & R. Luce (Eds.), *Contributions to the theory of games IV* (pp. 47-85). Princeton: Princeton University Press.

Gustafsson, J. E. (2015). Sequential dominance and the anti-aggregation principle. *Philosophical Studies, 172*(6), 1593-1601. doi:10.1007/s11098-014-0366-0

Gustafsson, J. E. (2020). Population axiology and the possibility of a fourth category of absolute value. *Economics and Philosophy, 36*(1), 81-110. doi:10.1017/S0266267119000087

Gustafsson, J. E., & Espinoza, N. (2010). Conflicting reasons in the small-improvement argument. *Philosophical Quarterly, 60*(241), 754-763. doi:10.1111/j.1467-9213.2009.648.x

Hewitt, E. (1960). The role of compactness in analysis. *The American Mathematical Monthly, 67*(6), 499-516. doi:10.2307/2309166

Holtug, N. (2001). On the value of coming into existence. *The Journal of Ethics, 5*(4), 361-384. doi:10.1023/A:1013957425591

Kreps, D. M. (2013). *Microeconomic foundations.* Princeton, NJ: Princeton University Press.

Kukushkin, N. S. (2008). Maximizing an interval order on compact subsets of its domain. *Mathematical Social Sciences, 56*(2), 195-206. doi:10.1016/j.mathsocsci.2008.01.003

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy, 61*, 343-377.

Luce, D. R., & Suppes, P. (1965). Preference, utility, and subjective probability. In D. R. Luce, R. R. Bush & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 250–410). New York: Wiley.

Machina, M. J. (1989). Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature, 27*(4), 1622-1668.

McClennen, E. F. (1990). *Rationality and dynamic choice: Foundational explorations*. Cambridge: Cambridge University Press.

McMahan, J. (1981). Problems of population theory. *Ethics, 92*(1), 96-127. doi:10.1086/292301

McMahan, J. (2009). Asymmetries in the morality of causing people to exist. In M. A. Roberts, & D. T. Wasserman (Eds.), *Harming future persons: Ethics, genetics and the nonidentity problem* (pp. 49-68). Dordrecht: Springer Netherlands.

Mendelson, B. (1962). *Introduction to topology*. Boston: Allyn and Bacon.

Mukherji, A. (1977). The existence of choice functions. *Econometrica, 45*(4), 889-894. doi:10.2307/1912679

Narveson, J. (1973). Moral problems of population. *Monist, 57*(1), 62-86. doi:10.5840/monist197357134

Nebel, J. M. (2019). An intrapersonal addition paradox. *Ethics, 129*(2), 309-343.

Ng, Y. (1989). What should we do about future generations? *Economics and Philosophy, 5*(2), 235-253. doi:10.1017/S0266267100002406

Parfit, D. (1987 [1984]). *Reasons and persons*. Revised ed. Oxford University Press.

Peijnenburg, Jeanne, & Atkinson, David. (2010). Lamps, cubes, balls and walls: Zeno problems and solutions. *Philosophical Studies, 150*(1), 49-59. doi:10.1007/s11098-009-9391-9

Priest, G. (1999). On a version of one of zeno's paradoxes. *Analysis, 59*(1), 1-2. doi:10.1111/1467-8284.00139

Pruss, A. R. (2018). *Infinity, causation, and paradox*. Oxford: Oxford University Press.

Pummer, T. (2019). The worseness of nonexistence. In E. Gamlund, & C. T. Solberg (Eds.), *Saving people from the harm of death* (pp. 215-228). New York: Oxford University Press.

Rabinowicz, W. (2009). Broome and the intuition of neutrality. *Philosophical Issues, 19*(1), 389-411. doi:10.1111/j.1533-6077.2009.00174.x

Rachels, S. (2001). A set of solutions to parfit's problems. *Noûs, 35*(2), 214-238. doi:10.1111/0029-4624.00294

Roberts, M. A. (2011). An asymmetry in the ethics of procreation. *Philosophy Compass, 6*(11), 765-776. doi:10.1111/j.1747-9991.2011.00435.x

Slote, M. (1989). *Beyond optimizing: A study of rational choice*. Harvard University Press.

Smith, T. (1974). On the existence of most-preferred alternatives. *International Economic Review, 15*(1), 184-194.

Sorensen, R. (1994). Infinite decision theory. In J. Jordan (Ed.), *Gambling on god: Essays on pascal's wager* (pp. 139-159). Lanham, Md.: Rowman & Littlefield.

Temkin, L. S. (1996). A continuum argument for intransitivity. *Philosophy and Public Affairs, 25*(3), 175-210. doi:10.1111/j.1088-4963.1996.tb00039.x

Temkin, L. S. (2012). *Rethinking the good: Moral ideals and the nature of practical reasoning*. New York; Oxford: Oxford University Press.

Thomas, T. (2018). Some possibilities in population axiology. *Mind, 127*(507), 807-832.

Varzi, A. (2015). Boundary. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (winter 2015 edition).

Weatherson, B. (ms). *Solving an infinite decision problem.* Unpublished manuscript. http://brian.weatherson.org/idt.pdf

Yablo, S. (2000). A reply to new zeno. *Analysis, 60*(266), 148-151. doi:10.1111/1467-8284.00217

Chapter 6

Transfinite Transitivity and Rational Choice

Abstract: This paper explores the relationship between rational choice theory and the principle of transfinite transitivity, a generalization of transitivity in contexts where our domain of evaluation is equipped with a notion of convergence. The interest of transfinite transitivity comes from its applications: in variable-population ethics it can be used to argue that it is always worse to create a miserable life and that it is always better to create a happy life, and in infinite-population ethics it can be used to establish the impossibility of combining a Pareto principle and a weak anonymity principle. This paper will argue that if transitivity can be supported by the role it plays in the theory of rational choice, then so can transfinite transitivity. In finite option sets, transitivity helps secure the existence of choiceworthy options, is implied by consistency conditions on permissible choice, and can be supported by money-pump arguments. This paper shows that transfinite transitivity plays the same role in infinite compact sets, and that nothing else can plausibly play the same role in infinite noncompact sets. The role of compactness is clarified and defended. The distinctions developed in the paper are then applied to a classic infinite-decision puzzle, Satan's Apple.[1]

Word count: 10040

---

[1] I would like to thank Michal Masny, Jens Jäger, Aidan Penn, and Todd Karhu. Special thanks to Ralf Bader, Teru Thomas, Theron Pummer, Tomi Francis, and Matthew Lau for helpful written comments.

1 Introduction

According to

> **Transitivity of Weak Worseness**. If x is at least as bad as y, and y is at least as bad as z, then x is at least as bad as z.

Put differently:

$$x \preccurlyeq y \preccurlyeq z \text{ implies } x \preccurlyeq z.^{1}$$

And according to

> **Acyclicity of Strict Worseness**. If x_1 is worse than x_2, x_2 is worse than x_3, ..., x_{n-1} is worse than x_n, then x_n is not worse than x_1.[2]

Put differently, the following pattern is prohibited:

$$x_1 \prec x_2 \prec \cdots \prec x_{n-1} \prec x_n \prec x_1.$$

Whenever our domain of evaluation has enough structure to allow for a natural notion of convergence, both principles can be generalized into the transfinite. To see how, start by chaining finitely many applications of transitivity together.

$$x_1 \preccurlyeq x_2 \preccurlyeq x_3 \text{ implies } x_1 \preccurlyeq x_3.$$

$$x_1 \preccurlyeq x_2 \preccurlyeq x_3 \preccurlyeq x_4 \text{ implies } x_1 \preccurlyeq x_4.$$

$$x_1 \preccurlyeq x_2 \preccurlyeq x_3 \preccurlyeq x_4 \preccurlyeq x_5 \text{ implies } x_1 \preccurlyeq x_5.$$

---

[1] I will write "$\preccurlyeq$" for "at least as bad as" (weak worseness), "$\prec$" for "worse than" (strict worseness), "$\succcurlyeq$" for "at least as good as" (weak betterness), "$\succ$" for "better than" (strict betterness) and "$\sim$" for "just as good as" (equal goodness). I take "$\preccurlyeq$" as basic, the rest is defined from it in the usual way.

[2] All conditions in this section can be adapted for binary relations in general such as strict betterness, equal goodness, or preference. When it does not matter, the binary relation involved will be left implicit.

And so on. We see that transitivity of weak worseness allows us to make shortcuts along *finite* worseness paths. Transfinite transitivity extends this pattern into the transfinite, by means of the notion of convergence, thus allowing us to make shortcuts along *infinite* worseness paths:

$$x_1 \preccurlyeq x_2 \preccurlyeq x_3 \preccurlyeq \cdots \rightarrow x \text{ implies } x_1 \preccurlyeq x,$$

where "$\rightarrow$" indicates that sequence $x_1, x_2, ...$, ordered by $\preccurlyeq$, converges to x.[3] Put differently:

> **Transfinite Transitivity of Weak Worseness**. If the sequence $x_1, x_2, ...$ converges to x, and x_1 is at least as bad as x_2, x_2 is at least as bad as x_3, ..., then x_1 is at least as bad as x.[4]

What do we mean by "convergence"? Perhaps the easiest way to understand it is in terms of *closeness*: the sequence $x_1, x_2, ...$ converges to x iff members of the sequence eventually become arbitrarily close to x.[5] For example, it makes sense to say that the sequence of numbers $1, \frac{1}{2}, \frac{1}{4}, ...$ converges to number 0, as the

---

[3] This notation is from Bartha, Barker & Hájek (2014) who also introduce the name "transfinite transitivity" for what I call "transfinite acyclicity of strict worseness". The idea of transfinite transitivity itself is apparently due to Gillies (1959). It was then reinvented by Smith (1974) and discussed by Birchenhall (1977), Mukherji (1977) and later by Carosi & Zaffaroni (1990) and Kukushkin (2008). It was reinvented again by both Bartha et al. (2014) and Weatherson (ms). Many of this paper's results are inspired by the economics strand of this literature.

[4] Two easy corollaries. First, since all subsequences of a convergent sequence converge to the same limit, transfinite transitivity implies that whenever $x_1 \preccurlyeq x_2 \preccurlyeq \cdots \rightarrow x$, then $x_i \preccurlyeq x$ for all i. Second, reflexivity plus transfinite transitivity implies transitivity: if $x \preccurlyeq y \preccurlyeq z$, then $x \preccurlyeq y \preccurlyeq z \preccurlyeq z \preccurlyeq \cdots \rightarrow z$, so, by transfinite transitivity, $x \preccurlyeq z$.

[5] This paper will assume that we have a quantitative measure of closeness, a *distance metric*. Consequently, all results hold for *metric spaces*, a subset of *topological spaces*. All mathematical concepts used in this paper are explained in standard topology textbooks, such as Kelley (1955).

number line comes equipped with a natural quantitative notion of closeness: absolute difference.

Analogously, we can define

> **Transfinite Acyclicity of Strict Worseness**. If the sequence $x_1, x_2, ...$ converges to x, and x_1 is worse than x_2, x_2 is worse than x_3, ..., then x is not worse than x_1.

The difference between transfinite transitivity and acyclicity concerns what happens in the limit as things get better and better: the former says that the limit is best, the latter that it is not worst. The two transfinite principles *look* like their finite namesakes. Of course, they differ in that they both presuppose some notion of convergence, unnecessary or absent in the finite case. Yet in many contexts we do have an intuitive grasp of convergence, so the transfinite principles can be formulated and assessed. The next section will give examples in variable-population ethics and infinite-population ethics. In both cases we will be able to use the transfinite principles to establish important new results.

But the main aim of this paper is to show that, in the theory of rational choice, transfinite transitivity plays an important role analogous to that of transitivity. Since many people are moved to endorse transitivity because of this role, they should arguably also be moved to endorse transfinite transitivity.[6] Because the two play analogous roles it seems the latter is plausibly a generalization of the former. After all, if something looks like transitivity, swims like transitivity, and quacks like transitivity, it is probably some kind of transitivity. But, to reiterate,

---

[6] Broome (2004: 50-63) has a different argument for transitivity, namely, that it is an analytic feature of relations expressed by comparatives. This paper's argument is different, although footnote 48 briefly discusses whether something like Broome's argument can support transfinite transitivity, too.

this paper's aim is conditional. Depending on what we think of transitivity, we can then either do a *modus tollens* or a *modus ponens*. Transfinite acyclicity will have an important place in this story, too, although the focus will be on transfinite transitivity.

Section 3 introduces the role of transitivity in guaranteeing the existence of choiceworthy options in finite sets, its relationship to consistency conditions relating permissible choices across finite option sets, and its relationship to money-pump arguments. Then there follow three sections, 4, 5, and 6, which take up the task of showing that transfinite transitivity can play these three roles role in a range of infinite option sets. These sections catalogue various results about transitivity and acyclicity and examine how far they can hold up in infinite decision problems. Section 7 concludes by applying the concepts developed in this paper to a classic infinite decision puzzle, Satan's Apple.[7]

[7] Another point in favour of transfinite transitivity is how it compares with principles of continuity which can play some (though not all) of transfinite transitivity's role in rational choice theory. See, for example, Uzawa (1956). Transfinite transitivity is notably weaker, allowing for lexicality and incomparability, two important possibilities ruled out by continuity. See Smith (1974), Mukherji (1977) and section 7 of Chapter 5. On incomparability and continuity, see Aumann (1962).

2 Two applications

Transfinite transitivity only becomes meaningful and interesting once combined with some appealing notion of convergence. Luckily, we can find plenty of such notions.[8]

2.1 Variable-population ethics

The first example concerns *variable-population ethics*, where we compare the goodness of worlds with different numbers of existing people.

For example, consider creating someone whose life which is *constantly bad*, in the sense that the longer they live, the worse-off they are overall. If we graph such a person's lifetime welfare (up to a given point) as a function of time, we get a downward-sloping curve. Its slope – extra lifetime welfare accumulated divided by time lived – is always negative.[9]

Now suppose that world w_1 is the result of adding to world w an extra person, call them "Zeno", with a constantly bad life of 80 years. Zeno accumulates –10 units of lifetime welfare in any given year. And let $w_2, w_3, ...$ be worlds where Zeno's lifespan is repeatedly cut in half. Zeno's life is not only bad but causally isolated. No one else is affected. This is shown in the following table.

[8] The two applications to follow are discussed in more detail in chapter 5, about variable-population ethics, and "Transfinite transitivity in infinite worlds" (unpublished), about infinite-population ethics. What follows is a mere sketch.

[9] Compare Broome (2004: 68) and Brown (ms).

	Zeno		Others
	Lifespan	Lifetime welfare	
w_1	80	–800	Unaffected
w_2	40	–400	Unaffected
…	…	…	…
w	Nonexistence		Unaffected

Table 6-1

It is plausible to think that the sequence of worlds $w_1, w_2, ...$ converges to world w. For example, the region of spatiotemporal difference between w and members of the sequence $w_1, w_2, ...$ is getting smaller and smaller. For i large enough, world w_i differs from world w only for a fraction of a second.

To make the claim of convergence even more plausible we can assume that Zeno's life is open on the left, so that there is no first moment of time when it is lived. Zeno's life is therefore like the left-open interval (0,1], as opposed to a left-closed interval like [0,1]. Hence, if Zeno's life were shortened to zero length, it arguably could not become a point-sized zero-length life, but would have to instead disappear altogether.

Now note that Zeno is worse-off in w_2 than in w_1, worse-off in w_3 than in w_2, and so on. Others are unaffected, so, arguably, they are equally well-off in each world. We conclude that each world in the sequence is worse than the next. This follows from

> **Strong Pareto**. If everyone is at least as well-off in world x as in world y, then x is at least as good as y, and if, in addition, some people are better off in x than in y, then x is better than y.

We therefore have the following pattern:

$$w_1 \prec w_2 \prec \cdots \rightarrow w.$$

Hence, by transfinite transitivity of strict worseness, it follows that

$$w_1 \prec w.$$

Since there is nothing special about Zeno, we conclude, in general, that creating a person with a constantly bad life is worse than not creating them. If so, then, arguably, creating any person with a negative-welfare life is worse than not creating them.[10] Hence, we obtain

> **The Negative Mere Addition Principle**. If worlds x and y differ only in that there is one extra person in y at a negative welfare level, then y is worse than x.

This principle is violated by many prominent theories in variable-population ethics: average utilitarianism, variable-value theories, and forms of egalitarianism.[11] Hence, we see that transfinite transitivity of strict worseness allows us to reach important conclusions on otherwise minimal assumptions, the chief one being about convergence.

---

[10] Transitivity and strong Pareto imply that if the argument works for Zeno's left-open and constantly bad life, it works for other lives with equal welfare. This makes it unlike some otherwise similar arguments in Bader's unpublished work. Another difference is that Bader nowhere appeals to transfinite transitivity and works directly with permissible choice rather than value. Nonetheless, this paper would not have existed without the inspiration of Bader's earlier work.

[11] See Arrhenius (2000). We can run an analogous argument on the positive side, this time using transfinite transitivity of *strict betterness*, leading to the much more controversial *positive* mere addition principle.

2.2 Infinite-population ethics

While notions of convergence have not been much used in variable-population ethics,[12] there are multiple well-studied notions of convergence in *infinite-population ethics*,[13] where we compare the goodness of worlds with infinitely many existing people.

For example, consider the following situation. We have a single indivisible dose of some magic potion which benefits those who drink it. And we have infinitely many people: One, Two, Three, and so on. Some of them would benefit more than others. One would get 2 units of welfare, Two would get 4 units, Three would get 8 units, and so on. Each of the infinitely many people involved can either drink the potion or pass it on to the next person in the line. This situation is depicted in the table below.

	People					
	One	Two	Three	Four	Five	
w_1	2	0	0	0	0	...
w_2	0	4	0	0	0	...
w_3	0	0	8	0	0	...
w_4	0	0	0	16	0	...
...						
w	0	0	0	0	0	...

Table 6-2

It is plausible to think that the sequence of worlds $w_1, w_2, ...$ converges to world w.

[12] Except in representation results such as Broome (2003) and Blackorby et al. (2001).

[13] See Campbell (1985) and Lauwers (1997).

This intuition seems to be shared by a number of philosophers who describe structurally similar cases. For example, Pollock (1983) imagines an immortal wine connoisseur in possession of a bottle of EverBetter Wine whose quality improves by the day. He suggests that if the connoisseur decides to wait every day, they never drink the wine. Same here: if the potion is always passed on, no one drinks it.

In infinite-population ethics, this verdict is captured by a notion of convergence which we can call *pointwise convergence*.[14] Put roughly, over the sequence $w_1, w_2, ...$ each person's welfare converges to their welfare in w.[15] Indeed, for every person there is a point in the sequence after which their welfare is zero, the same as in w.

It is moreover plausible that each world in the sequence $w_1, w_2, ..$ is worse than the next. Given ordinary transitivity, this follows from the conjunction of strong Pareto and

> **Finite Anonymity**. Worlds x and y are equally good if y differs from x only in that welfare levels of finitely many people are interchanged.

We therefore have the following pattern:

$$w_1 \prec w_2 \prec \cdots \rightarrow w.$$

Hence, by transfinite transitivity of strict worseness:

$$w_1 \prec w.$$

---

[14] This corresponds to convergence in the product topology, see Diamond (1965).

[15] Put more precisely: for each person and any positive degree of closeness ϵ, there is a point in the sequence $w_1, w_2, ...$ after which their welfare is within ϵ of their welfare in w.

But this contradicts strong Pareto which implies that w_1 is better than w, a world where everyone has zero welfare. Hence, using transfinite transitivity of strict worseness, we can show the impossibility of combining strong Pareto, finite anonymity, ordinary transitivity and the pointwise notion of convergence. This is an important result, contributing to many extant impossibility results in infinite-population ethics.[16]

[16] See van Liedekerke's (1995) impossibility result which shows that a stronger anonymity principle is incompatible with strong Pareto, and Diamond's (1965) first theorem which shows that finite anonymity, strong Pareto and continuity are incompatible, given pointwise convergence. The argument in the main text can also be adapted for weaker notions of convergence (for example, uniform convergence). See my "Transfinite transitivity in infinite worlds" (unpublished).

3 Compactness in the theory of rational choice

The last section argued that transfinite transitivity and, by extension, acyclicity are worth studying. The rest of the paper will show what they can do for us in rational choice theory.

In a very abstract sense, rational choice is about selecting options from arbitrary option sets, without presupposing much about their inner structure. The selection process is supposed to be rational in some sense, perhaps guided by a binary relation defined on the set. We would therefore like to know: (1) Given an option set what should one select from it? (2) How do permissible selections from different sets relate to each other? (3) And how do permissible selections at different points in time relate to each other?

Transitivity and acyclicity play an important role in answering these questions insofar as finite option sets are concerned. First, they help ensure the existence of choiceworthy options. Second, they are intimately related to standard consistency conditions on choice which constrain how choices are to be made across sets and their subsets. Third, their violators are liable to be exploited in money pumps, where sequences of trades are offered across time.

With infinite option sets, the situation is more complicated. Roughly speaking, there appear to be tame and wild infinite sets. Suppose, for example, the more money the better and consider $\{\$1, \$2, \$3, ...\}$, the set of natural-valued dollar amounts, or $[\$0, \$100)$, the set of dollar amounts no less than \$0 but strictly less than \$100. In either case no option is *maximal* in the sense of not being worse than any available option. *A fortiori*, no option is *best* in the sense of being at least as good as any available option. But it is often assumed that, setting deontological considerations aside, one is permitted to choose all and only maximal options. So, it is puzzling what one is to choose in these cases.

Here the puzzlement arguably arises because the set of available options, A, fails to have the property of

> **Compactness**. Every sequence drawn from A has a subsequence which converges to something in A.

In our examples, the sequence $\$1, \$2, \$3, ...$ diverges to infinity, so has no convergent subsequence, and even though the sequence \$0, \$50, \$75, \$92.5, … converges to \$100, \$100 itself is not an available option. I will call situations like these *EverBetter Problems*.

For EverBetter Problems to be possible it is enough to assume the innocuous condition of

> **Monotonicity**. There is some quantity Q such that, other things being equal, the more of Q the better, and Q is either unbounded above or continuous.[17]

This condition's second clause is meant to ensure that if Q is capped above, then we can still make sense of approaching that cap arbitrarily closely from below.

Now consider $\{\$1, \$2, \$3, ..., \$ + \infty\}$, the set of natural-valued dollar amounts on the extended number line, with "positive infinity" thrown in, or $[\$0, \$100]$, the set of dollar amounts no less than \$0 and no more than \$100. These sets can be seen as "compactified" versions of the two option sets considered before. And it is clear what the choice from them should be: $\$ + \infty$ and \$100, respectively. Hence, compactness seems to divide unproblematic infinite decision situations from the

[17] "Unbounded above" does not mean that one's von Neumann-Morgenstern utility function for money, say, is unbounded above. It is the quantity Q that is unbounded above, not the utility of Q.

potentially problematic ones. We should expect rational choice to be possible from infinite compact sets but not necessarily from infinite noncompact ones.[18]

A competing account is that EverBetter Problems are simply counterexamples to standard approaches to rational choice, such as *maximizing choice*, according to which all and only maximal options are permissible, and *optimizing choice*, according to which all and only best options are permissible. Both approaches imply that EverBetter Problems are rational dilemmas: no option is permissible.

It might therefore seem that the right lesson to draw is instead to reject maximizing and optimizing choice in favour of *satisficing choice*, according to which the permissible options are all and only those which are *good enough* in the relevant context. This was Slote's (1989) conclusion.[19] In this light, compactness might not seem important anymore.

But if, following Hurka (1990), we make a distinction between two kinds of satisficing, we will see that satisficers have their own EverBetter Problems. According to *absolute satisficing*, the threshold of good enough is selected independently of the option set facing the agent and "[w]hen a situation is and will remain below the absolute threshold, an agent's duty is the same as under maximizing: she must do everything to move it towards satisfactory goodness" (107–108). According to *comparative satisficing*, the threshold of good enough depends on the option set, so that "an agent's duty is always but only to bring

---

[18] Compactness is routinely assumed in economics in this context, see Kreps (2013: 1-29). It might actually be too strong: it can be weakened along the lines suggested by Carosi & Zaffaroni (1990: 280-282).

[19] Arntzenius et al. (2004), Meacham (2010), and Bartha et al. (2014) also suggest satisficing responses to EverBetter Problems.

about some reasonable percentage of the largest contribution to goodness she can" (108).

It is now easy to see that absolute satisficers face EverBetter Problems below the absolute threshold. For example, assuming the threshold is set at \$0, consider the problem of choosing something from $[\$-100, \$0)$, the set of dollar amounts no less than \$–100 and strictly less than \$0. This problem stumps absolute satisficers just as much as it does maximizers.[20]

On the other hand, comparative satisficers are stumped in EverBetter Problems where there is no such thing as the best, like in the problem of choosing something from $\{\$1, \$2, \$3, ...\}$. They might be stumped even if the option set is bounded above like in the problem of choosing something from $[\$0, \$100)$. If goodness lacks sufficient quantitative structure, so that it is more like temperature rather than mass, the instruction to do something at least 80%, say, as good as the best will be meaningless.[21] Comparative satisficing in Hurka's sense becomes impossible. Hence, it is not just maximizers and optimizers but also satisficers who have a reason to care about compactness.

The last point in favour of the importance of compactness is that, in the theory of rational choice, its role is similar to that in mathematics at large which Hewitt (1960) describes as follows:

> "Compactness of subsets of $\mathbb{R}$ is a generalization of finiteness. (…) *a great many propositions of analysis are*: (A) *trivial for finite sets*; (B) *true and reasonably simple for infinite compact sets*; (C) *either false*

[20] Bradley (2006) also makes this point, crediting it to Gustaf Arrhenius.

[21] For example, the instruction to set one's thermostat to 30% of water's boiling temperature is meaningless. That temperature can be either 30°C or 63.6°F (which is about 17.5°C), depending on an arbitrary choice between two temperature scales.

> *or extremely difficult to prove for noncompact sets.* (…) some care is usually needed in moving from the finite situation to the corresponding infinite compact situation. A given assertion true for finite sets often needs some qualification to be provable for all compact sets" (500).

Each of the following three sections will likewise break down into three parts: we will start with easy results about transitivity and acyclicity in finite sets, generalize using transfinite transitivity and acyclicity to infinite compact sets, and then show how they break down in infinite noncompact sets.[22]

[22] Note that finite sets (if equipped with a topology) are automatically compact. Hence, after covering finite option sets we can go straight to infinite compact ones.

4 Existence of choiceworthy options

As we saw, it is often assumed that choiceworthy options in A are just the best options in A, or at least just the maximal options in A. In finite option sets, transitivity and acyclicity have a role in guaranteeing the existence of choiceworthy options in that sense, as shown by the following standard results.[23] The first is:

(1) The existence of maximal options in all finite option sets implies acyclicity of strict worseness.

The converse is also true:

(2) Acyclicity of strict worseness implies the existence of maximal options in all finite option sets.

Transitivity of weak worseness, on the other hand, has a role in securing the existence of best options:

(3) Completeness and transitivity of weak worseness imply the existence of best options in all finite option sets.

The first and last of these carry over when "finite" is replaced by "infinite compact" and "transitivity" and "acyclicity" are prefixed with "transfinite". Highlighting the changes, the first becomes:

(4) The existence of maximal options in all **infinite compact** option sets implies **transfinite** acyclicity of strict worseness.

To see this, let $A = \{x, x_1, x_2, \dots\}$ and suppose that $x \prec x_1 \prec \cdots \to x$, a transfinite cycle of strict worseness. Note that A is compact, as every sequence drawn from

[23] For proofs, see Sen (1970: 1-20).

A either has a subsequence which is eventually constant (like $x_1, x_1, ...$) or shares a subsequence with $x_1, x_2, ...$ (like $x_1, x_3, x_5, ...$), so, either way, it has a convergent subsequence. But no option in A is maximal. Hence, transfinite acyclicity of strict worseness is necessary for the existence of maximal options.[24]

Perhaps surprisingly there is a difference between transfinite cycles of *strict betterness* and *strict worseness*. The absence of the former is not necessary for the existence of maximal options. To see this, let A be as before but suppose that $x \succ x_1 \succ x_2 \succ \cdots \to x$, a transfinite cycle of strict betterness. Again, A is compact. But it has a maximal option, namely, x. What this example also shows is that while transfinite cycles of strict betterness are compatible with the existence of maximal options, they are incompatible with the existence of *minimal* options, defined as not better than any available option. Since in rational choice theory we are more interested in maximal options, in the infinite case strict betterness turns out to be less important than strict worseness. No similar difference arises in the finite case.[25]

The middle result from the finite case does not carry over to the infinite compact case:

(5) **Transfinite** acyclicity of strict worseness **does not imply** the existence of maximal options in all **infinite compact** option sets.[26]

---

[24] Essentially the same argument can be found in Birchenhall (1977), Mukherji (1977), and Kukushkin (2008).

[25] This difference between transfinite cycles of strict worseness and strict betterness matters for variable-population ethics, see chapter 5.

[26] Birchenhall (1977) has a similar example, crediting it to Ted Bergstrom. The example to follow is essentially Kukushkin's (2008), as is the diagnosis of the problem.

It is instructive to see why transfinite acyclicity of strict worseness is too weak. Take an analog clock like the one below.

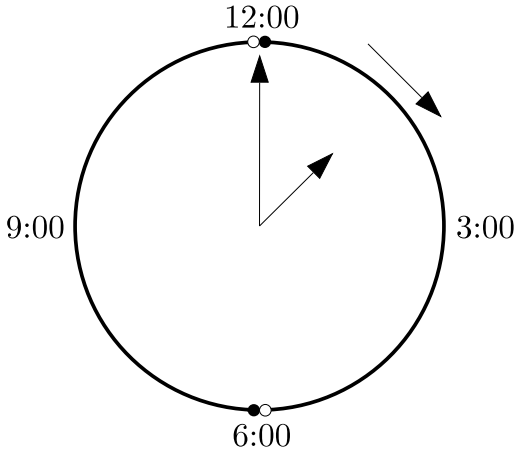


Figure 6-1. Kukushkin's clock

Divide its perimeter in two halves: the right half including 12:00 but excluding 6:00, and the left half composed of the rest. Order times in each half as usual but declare times from different halves as incomparable. Convergence is defined based on the natural notion of distance, measured along the perimeter.

We can then see that the relation we defined is transfinitely acyclic. For example, each time in the sequence 12:00, 3:00, 4:30, 5:45, … is ranked below the next, with the sequence itself converging to 6:00. So, none of 12:00, 3:00, 4:30, … are maximal. Yet 6:00 is not maximal either, as it is ranked below 9:00, for example. So, no time is maximal with respect to our defined relation. Yet the set of times marked off on the clock is compact.

The problem is that transfinite acyclicity only prohibits cycles which can be indexed by the first transfinite ordinal, ω. The example with the clock does contain a cycle of sorts, but one which can only be indexed by a bigger transfinite ordinal $\omega + \omega$. To see this, start at 12:00, make ω-many improvements to end up at 6:00. Then move to the better 9:00. This is the $(\omega + 1)$-th move. Then move

to the even better 10:30. And so on. Hence, after $\omega + \omega$-many improvements we go back to 12:00.[27]

By contrast, the third result from the finite case does carry over:

(6) Completeness and **transfinite** transitivity imply the existence of best options in all **infinite compact** option sets.[28]

Some intuition for this result can be gained by considering the special case of countably infinite option sets. If a set A is countably infinite, it can be enumerated like so: $x_1, x_2, x_3, ...$, where every member of A appears somewhere on this list (perhaps more than once). We can therefore think of A as a sequence. If A has no best elements, we can find an infinite subsequence $y_1, y_2, y_3, ...$, where each member is worse than the next. Simply start at x_1 and keep going to the next member in the enumeration which is better than the last member selected. By compactness, $y_1, y_2, y_3, ...$ has a subsequence $z_1, z_2, z_3, ...$, converging to some z^* in A, where each member is also worse than the next. It is then easy to see that, by transfinite transitivity of strict worseness, every z in $z_1, z_2, z_3, ...$ is worse than z^*. Now note that, by completeness, if some x from $x_1, x_2, ...$ is not selected into the $y_1, y_2, ...$ subsequence, it must be at least as bad as some y in that subsequence.

---

[27] Prohibiting cycles indexed by higher (countable) ordinals is not enough either. We can use another "unit-circle" clock to show this. Fix Δ to be some irrational multiple of 2π (the clock's circumference). Rank point y over point x if the former can be reached from the latter by some number of moves of length Δ clockwise along the perimeter, and otherwise declare y and x incomparable. The defined relation has no cycles of length $\omega, \omega + \omega, \omega + \omega + \omega$, and so on, the set of points on the unit circle is compact, yet there is no maximal point with respect to the defined relation. This is my rendering of Kukushkin's (2008) example. See also Devaney (1989: 21-22). How to strengthen transfinite acyclicity to deal with Kukushkin's two counterexamples appears to be an open question.

[28] This follows from part (b) of Theorem 4.1 in Smith (1974).

And, likewise, by transitivity, every y in the $y_1, y_2, ...$ subsequence is at least as bad as some z in the $z_1, z_2, z_3, ...$ subsequence. Hence, by transitivity, every x from $x_1, x_2, x_3, ...$ is worse than z^*. So, z^* is best in A after all.

This argument will not necessarily work if the option set A is uncountably infinite, since then it cannot be enumerated and cannot be thought of as a sequence. But the result itself can be generalized, as shown by Smith (1974).[29]

Note that the argument we just went through (and Smith's general one) only uses transfinite transitivity of strict worseness in addition to transitivity and completeness. So, we can conclude:

(7) For complete preorders, **transfinite** transitivity of strict worseness is sufficient for the existence of best options in all **infinite compact** sets,

where a *complete preorder* is a relation which is complete in addition to being reflexive and transitive.

And, luckily, for complete preorders, transfinite transitivity of strict worseness is also necessary for the existence of best options in all infinite compact sets.[30] To see this, let $A = \{x, x_1, x_2, ...\}$, suppose that A has a best option and that $x_1 \prec x_2 \prec \cdots \to x$. Then since none of $x_1, x_2, ...$ can be best in A, x must be best. But it cannot be that, for some i, $x_i \sim x$. That would mean $x \sim x_i \prec x_{i+1}$. Hence, by transitivity, $x \prec x_{i+1}$, contradicting the claim that $x_{i+1} \preccurlyeq x$, which follows if x is

---

[29] Smith's (1974) proof uses the fact that compactness implies the finite intersection property. By contrast, my proof can be generalized to uncountable compact sets by using nets (which generalize sequences) and reformulating compactness to require that every net has a convergent subnet, making the condition of compactness more closely related to the demand that EverBetter Problems be avoided. On nets, see Kelley (1955: 62-83).

[30] This is essentially Smith's (1974) argument in part (a) of Theorem 4.1.

best. So, it must be that $x_i \prec x$ for all i, as per transfinite transitivity of strict worseness. So, putting the two things together:

(8) For complete preorders, **transfinite** transitivity of strict worseness is equivalent to the existence of best options in all **infinite compact** sets.[31]

Lastly, we get to noncompact option sets. It is easy to see that the previous claims of necessity carry over, while those of sufficiency fail. In particular,

(9) For complete preorders, **transfinite** transitivity of strict worseness is **not** sufficient for the existence of best options in all **infinite** sets.

Hence, when we remove "compact" before "infinite", we lose our key result. To see this, first record that

(10) The "greater than or equal to" ($\leq$) and the "greater than" ($<$) relations are transfinitely transitive on the number line, $\mathbb{R}$, with its usual topology.

The algebraic manipulations needed to show this are relegated to a footnote.[32] Then take an EverBetter Problem. If, as per monotonicity, value tracks the $\leq$ relation on some real-valued quantity Q (such as money), then transfinite

---

[31] This is Smith's (1974) Theorem 4.1. It can be weakened by dropping "complete" provided that "best" is changed to "maximal". See Birchenhall (1977) and Kukushkin (2008). They also show that transfinite transitivity of strict worseness can be weakened to the condition that the strict worseness relation can be extended in a way which makes it transfinitely transitive.

[32] Let $r_1, r_2, ...$ be real numbers converging to r and let $r_i \leq r_{i+1}$ for all $i \geq 1$. Suppose that, contrary to transfinite transitivity of $\leq$, $r_j \not\leq r$ for some j. Then, since $\leq$ is complete, $r_j > r$ for that j. Hence, $r_j = r + \delta$, with $\delta > 0$. Since $r_i \leq r_{i+1}$ for all $i \geq 1$, it follows, by transitivity of $\leq$, that $r + \delta \leq r_i$ for all $i \geq j$. Hence, it is not true that for any $\epsilon > 0$ the distance between r and members of the sequence $r_1, r_2, ...$ is eventually less than ϵ, since it cannot be eventually less than $\delta > 0$. This contradicts the claim that the sequence $r_1, r_2, ...$ converges to r. A similar argument shows that $<$ is transfinitely transitive on the number line with its usual topology.

transitivity of weak and strict worseness, completeness, and transitivity are automatically satisfied. Yet, as we saw in section 4, if the option set is noncompact, maxima and optima may be missing, as in the problem of choosing something from $[\$0, \$100)$. Hence, transfinite transitivity principles are powerless to guarantee the existence of best options in noncompact sets, if we insist on assuming monotonicity.

5 Internal consistency of choice

We saw that, like transitivity in finite option sets, transfinite transitivity helps ensure the existence of best options in infinite compact option sets. Transfinite acyclicity is less helpful in this regard, as it is too weak. We will now consider another role that transitivity and acyclicity can play and which transfinite transitivity and acyclicity can take up in infinite compact option sets.

Choice consistency conditions constrain how permissible choices relate across different option sets. They are not primarily about how choice relates to value (the topic of previous section) but how choices from different option sets relate to each other. Still, I will focus on agents who choose what is best at least sometimes. In particular, I will assume

> **Pairwise Guidance**. Only y may be chosen iff x is worse than y, provided that x and y are the only options available.

This is a weak connection between value and permissible choice: nothing yet follows about permissible choices from triples, quadruples, or indeed infinite sets. We can add choice consistency conditions to establish a stronger connection. Two are central, both due to Sen (1969), the first being

> **Property $\boldsymbol{\alpha}$**. If x may be chosen from A and x belongs to B, a subset of A, then x may be chosen from B.

The second is

> **Property $\boldsymbol{\beta}$**. If x and y both belong to A, a subset of B, and x and y may be chosen from A, then either both or neither may be chosen from B.

Properties α and β imply that permissible choice is context-free in some sense. The former is known as contraction consistency, while the latter is known as

expansion consistency. Another common assumption, made by Kreps (2013: 3), among others, is

> **Finite Nonemptiness**. If A is finite, then something may be chosen from A.

The thought is that while infinite option sets might be wild, finite option sets are always tame enough to allow for *some* permissible choice.

Then the connection between transitivity and acyclicity and rational choice can be summarized by two results, the first being:

(11) Pairwise guidance, property α, and finite nonemptiness imply acyclicity of strict worseness.

To obtain transitivity of strict worseness, we need to add property β into the mix:

(12) Pairwise guidance, property α, property β and finite nonemptiness imply transitivity of strict worseness.[33]

These results carry over with suitable changes to infinite compact sets.[34] We first need to define the analogue of finite nonemptiness, namely,

> **Compact Nonemptiness**. If A is **compact**, then something may be chosen from A.

Then the first result above corresponds to:

(13) Pairwise guidance, property α, and **compact** nonemptiness imply **transfinite** acyclicity.

---

[33] The proofs are variations on standard results, to be found in Sen (1969, 1970: 1-20 1971), and, so, are omitted.

[34] Mukherji (1977) and Kukushkin (2008) report related results.

To see this, let $A = \{x, x_1, x_2, ...\}$ and suppose that $x \prec x_1 \prec x_2 \prec \cdots \to x$. First, suppose that some option of the form x_i may be chosen from A. Then, by property α, it follows that it may be chosen from $\{x_i, x_{i+1}\}$, a subset of A. But, by pairwise guidance, only x_{i+1} may be chosen from that subset. Hence, nothing of the form x_i may be chosen from A. By a similar argument, x cannot be chosen from A either. So, nothing may be chosen from A. Since A is compact, this contradicts compact nonemptiness. Hence, our supposition that A contains a transfinite cycle must have been wrong and, so, transfinite acyclicity follows.

The second result becomes:

(14) Pairwise guidance, property α, property β, and **compact** nonemptiness imply **transfinite** transitivity of strict worseness.

Let $A = \{x, x_1, x_2, ...\}$ and suppose $x_1 \prec x_2 \prec \cdots \to x$. By a similar argument as above, property α implies that no option of the form x_i may be chosen from A. Since A is compact, compact nonemptiness implies that x is the sole permissible option in A. By property α again, x may be chosen from all subsets of the form $\{x, x_i\}$. If x_i also were permissible in that subset, then, by property β, it would be permissible in A, too. But it is not. Hence, x is the sole permissible option in all subsets of the form $\{x, x_i\}$. But, by pairwise guidance, this must be because all options of the form x_i are worse than x, as per transfinite transitivity of strict worseness.

At last, we get to noncompact sets, where things break down. We first define

> **General Nonemptiness**. If A is an option set, then something may be chosen from A.

Note that no restrictions are imposed on A, so that it may be noncompact.

We saw that, in finite option sets, and in the presence of pairwise guidance and finite nonemptiness, property α rules out cyclicity of strict worseness, while

adding property β rules out intransitivity of strict worseness. In compact sets, similarly, property α rules out **transfinite** cyclicity of strict worseness while adding property β rules out **transfinite** intransitivity of strict worseness. In noncompact sets, however, property α, together with the auxiliaries, might rule out too much:

(15) Pairwise guidance, property α, and **general** nonemptiness are **incompatible with** monotonicity.

Recall that according to monotonicity there is a quantity Q whose increases are always good and which is either unbounded above or continuous. The trouble is once again with EverBetter Problems. Suppose one has to choose something from $A = \{\$1, \$2, \$3, ...\}$. And suppose it is permissible to choose $\$n$. Then, by property α, $\$n$ can also be chosen from $\{\$n, \$(n+1)\}$, a subset of A. But given pairwise guidance, only $\$(n+1)$ is a permissible option in that set. After all, we are assuming the more money the better. This argument works for all options in A. Hence, no option may be chosen from A. But this contradicts general nonemptiness.

This is essentially Sorensen's (1994) argument against Pollock's (1983) suggestion that in EverBetter Problems anything is permissible:

> "If the transition to infinite choice really washed out the differences, then one could rationalize finite decisions by adding infinitely many more. (…) These extra alternatives do not make the choice of \$1 as rational as the choice of \$2" (147).[35]

While, as pointed out by Sen (1993: 501), among others, there are cases in which property α seems dubious, this does not seem to be one of these cases. And

[35] Kreps (2013: section 1.6) offers a similar argument.

monotonicity seems innocuous, too. Again, it is noncompactness itself that seems troublesome. If we want to hold on to monotonicity, pairwise guidance, and property α, we have to reject general nonemptiness, and think of these EverBetter Problems as rational dilemmas. We can bolster this diagnosis by noting that:

(16) Pairwise guidance, property α, and **compact** nonemptiness are **compatible** with monotonicity.

To see this, first note that monotonicity is compatible with completeness, transitivity, and transfinite transitivity of strict worseness, and, second, that these three conditions are sufficient for the existence of best options in all compact sets. It is then easy to verify that the choice method of always choosing the best options with respect to the relevant relation (if they exist) satisfies pairwise, property α, and compact nonemptiness.

6 Money pumps

Besides lack of choiceworthy options and violations of choice consistency conditions, another problem that might arise unless transitivity and acyclicity are in place is liability to *money pumps*.

6.1 Finite pumps

Two sorts of money pumps will be discussed. In line with tradition, I present the first sort of pump as directed against cyclic *strict dispreference* rather than *strict worseness*. To see how it works, consider Tversky's (1969) description:[36]

> "Suppose an individual prefers y to x, z to y, and x to z. It is reasonable to assume that he is willing to pay a sum of money to replace x by y. Similarly, he should be willing to pay some amount of money to replace y by z and still a third amount of money to replace z by x. Thus, he ends up with the alternative he started with but with less money" (45).

So, in addition to a cycle in the agent's ranking, Tversky's pump also assumes

> **Payment Continuity**. If x is preferred to y, then $x - \epsilon$ is preferred to y, for all sufficiently small positive ϵ,

where $x - \epsilon$ is an option just like x except worse by some small amount ϵ of money or other relevant currency. In what follows we will also apply this principle with "better than" replacing "preferred to".

[36] The original money pump is due to Davidson et al. (1955).

Another sort of money pump makes no reference to initial endowments nor does it require the agent to assess each trade in isolation. Instead, we assume that the agent conforms to

> **Naïve Choice**. An agent is permitted to choose a given action iff that action is part of a plan whose outcome can be permissibly chosen from the set of outcomes of all plans available at the time of action.[37]

Such a *naïve agent* is minimally strategic in a way in which Tversky's agent is not. If they trade y for z, this is not just because they prefer y to z but because that trade is called for by a plan they consider acceptable for the decision problem as a whole. Tversky's agent is *myopic*: not looking beyond the next trade.[38]

Naïve agents guided by cyclic or intransitive rankings can also be pumped, however. To see this, consider the following decision tree, due to Cantwell (2003: 389), where squares represent occasions of choice.

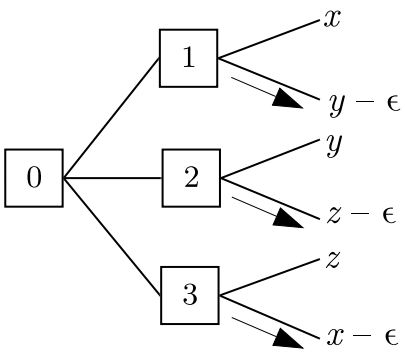


Figure 6-2. Cantwell's pump

In this example, we let the agent pick something among x, y, z, and, after they choose, we let them reconsider for a small payment. Assume, as in Tversky's

[37] See McClennen (1990). This formulation is close to Buchak's (2013: 175-6).

[38] The naïve/myopic distinction is explained in Buchak (2013: 219).

example, that $x \prec y \prec z \prec x$. By payment-continuity, it also follows that $x \prec y - \epsilon$, $y \prec z - \epsilon$, $z \prec x - \epsilon$ for some positive ϵ.

If a naïve agent reaches any of the nodes following the initial one, they will go down (as indicated by the arrows). At node 1, for example, only x and $y - \epsilon$ are available. Since x is worse than $y - \epsilon$, it follows, by pairwise guidance, that only the latter is permissible. Hence, whatever our naïve agent does at the initial node, they will subsequently go down. And since the set of outcomes reachable from the initial node is finite, finite nonemptiness implies that a naïve agent will do something at that node. Hence, in two moves, a naïve agent would end up paying for something they could have had for free.[39]

In Cantwell's pump that is what the problem is supposed to be.[40] By contrast, in Tversky's pump, the problem is supposed to be more specific: the agent ends up with what they had at the beginning except for less money.

I will say that an agent is liable to a *forcing money pump* if, given the ranking that guides them, they ought to act so that they pay for something they could have had for free. And a *non-forcing money pump* if, given that ranking, they are merely permitted to act in that way.[41] Hence:

---

[39] There are tricks to protect oneself against money pumps even if one's ranking is cyclic or intransitive. The key is to reject naïve choice. The main alternatives are *sophisticated choice* and *resolute choice*, although sophisticated choice does not help with Cantwell's pump. See McClennen (1990) and Buchak (2013: 170-200). There is no need to argue against resolute choice, however, since the aim of this paper is conditional: if money-pump arguments can support transitivity, they can also support *transfinite* transitivity.

[40] See Cantwell (2003). An alternative diagnosis appeals to the idea that $x - \epsilon$ is *covered* by x in the sense of Miller (1980).

[41] The forcing/non-forcing distinction appears in Gustafsson & Espinoza (2010).

(17) Cyclicity of strict worseness implies liability to forcing money pumps, assuming naïve choice, payment continuity, finite nonemptiness, and pairwise guidance.

Intransitivities can also make one liable to money pumps, albeit non-forcing ones. Assuming pairwise guidance, the agent will be at best merely permitted (not required) to go down at the subsequent nodes. Hence, a naïve agent might still go down. So:

(18) Intransitivity of weak worseness implies liability to non-forcing money pumps, assuming naïve choice, payment continuity, finite nonemptiness, and pairwise guidance.

6.2 Infinite pumps

Both Tversky's and Cantwell's pumps can be made to work in the infinite compact case, although the latter carries over more smoothly, with fewer extra assumptions.

We start with a version of Tversky's pump. Suppose we have a transfinite cycle of strict dispreference: $x \prec x_1 \prec x_2 \prec \cdots \to x$. The focus on dispreference rather than worseness is to keep in line with Tversky's earlier description.

We make two extra assumptions. First, that x is the initial default: if nothing is done, x will be the outcome. And that $x_1, x_2, ...$ can be seen as the results of making successive changes to that default. Then begin by offering the agent to change from x to x_1. Assuming they assess each trade in isolation, they will accept. Then offer them to change that to x_2. Once again, they will accept. And so on. Speed up as you go, so that an infinity of offers is made and accepted in finite time.

This is a *supertask*: a situation where infinitely many actions are performed in finite time. If we could argue that the end-state of this supertask is outcome x again, we could easily pump the agent: just ask them to pay a little to make the first trade.

But supertasks are controversial. There are some supertask stories where the end-state of a supertask appears underdetermined, as in the story of Thomson's (1954) lamp. The lamp is switched on at 12:00, then off at 12:30, then on at 12:45, then off at 12:52:30, and so on. What is the lamp's state at 13:00? Is it on or off? Grünbaum (1970) and Earman & Norton (1996) argued that different ways of implementing the lamp story lead to different end-states. For example, if the lamp is switched on by a bouncing ball closing an electric circuit and the infinite number of switches is accomplished by making the ball bounce back to ever lower heights, it plausibly follows that the end-state of the lamp is on.

In general, it seems that the end-state of a supertask is to be found by continuously extrapolating the outcome at its finite stages, so that the end-state is where the sequence of before-states converges.[42] How is this convergence ("supertask convergence") related to the sort of convergence employed by transfinite transitivity and acyclicity ("convergence")?

If convergence implies supertask convergence, then violations of transfinite acyclicity, for example, would make one liable to a pump as above: one could construct a situation where x is the default, make the agent go through an infinity of individually favourable trades, perhaps charging a little, and get them back to x. On the other hand, if supertask convergence implies convergence, then transfinite acyclicity, for example, would offer some degree of protection against

[42] See, for example, Allis & Koetsier (1995), Earman & Norton (1996), and Laraudogoitia (2016)

such pumps, since if one ever found oneself in a supertask scenario like the one described above, that supertask's end-state would have to be no worse than the status quo.

But note that the principles of transfinite transitivity and acyclicity themselves make no reference to supertasks, only to convergence. And, as we just saw, to defend them by pumps like Tversky's calls for lots of extra assumptions whose status is not perfectly clear. By contrast, Cantwell's pump can be adapted much more easily to the infinite compact case, as follows.

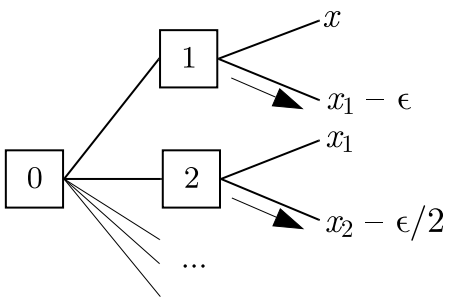


Figure 6-3. Infinite Cantwell pump

Here we offer the agent to pick something amongst $x, x_1, x_2, x_3, ...$, and, after they choose, we offer them a chance to reconsider for a sufficiently small payment.

This infinite Cantwell pump requires no supertasks to operate. It merely requires that the agent can be faced with a countable infinity of actions at once. The tree here is infinite in width but finite in length.[43]

To see how the pump works, let $A = \{x, x_1, x_1 - \epsilon, x_2, x_2 - \frac{\epsilon}{2}, x_3, ...\}$, the set of all outcomes available in this decision tree, and suppose that we have a transfinite cycle of strict worseness: $x \prec x_1 \prec x_2 \prec \cdots \rightarrow x$. Payment-continuity implies that

[43] For example, Pruss (2018: 106-111), a recent supertask sceptic, would presumably reject the possibility of the previous pump, but not this one.

no generality is lost by assuming that $x \prec x_1 - \epsilon, x_1 \prec x_2 - \frac{\epsilon}{2}$, and so on. And, plausibly, if the sequence $x_1, x_2, ...$ converges to x, then so does the sequence $x_1 - \epsilon, x_2 - \frac{\epsilon}{2}, x_3 - \frac{\epsilon}{4}, ...$, since the "main" outcomes converge to x and the "side payments" converge to 0. It follows from all this that A is compact.

At the initial node the agent has to choose between future choices. At the subsequent nodes they have to choose from pairs of options. And since $x \prec x_1 - \epsilon, x_1 \prec x_2 - \frac{\epsilon}{2}$, and so on, it follows, by pairwise guidance, that at all these subsequent nodes going down is required. Since A, the set of outcomes reachable from the initial node is compact, it follows, by compact nonemptiness and naïve choice, that the agent is permitted to do something at that node.[44] Hence, in two moves, a naïve agent will end up paying for something free.

This means that the first result about Cantwell's finite pump carries over to the infinite compact case:

(19) **Transfinite** cyclicity of strict worseness implies liability to forcing money pumps, assuming naïve choice, payment-continuity, **compact** nonemptiness, and pairwise guidance.

As does the other result:

(20) **Transfinite** intransitivity of weak worseness implies liability to non-forcing money pumps, assuming naïve choice, payment-continuity, **compact** nonemptiness, and pairwise guidance.

Things again break down in the noncompact case where EverBetter Dilemmas come back to bite us:

[44] If we did not assume compact nonemptiness, we could not conclude that.

(21) **Monotonicity** implies liability to forcing money pumps, assuming naïve choice, payment-continuity, **general** nonemptiness, and pairwise guidance.

Suppose, for example, the agent is asked to choose the number of days they will spend in Paradise. After they choose, they will be given a chance to switch to the next higher number. This is illustrated in the following version of the infinite Cantwell pump.[45] Then, after they choose, give the agent a chance to reconsider and switch to the next higher number.

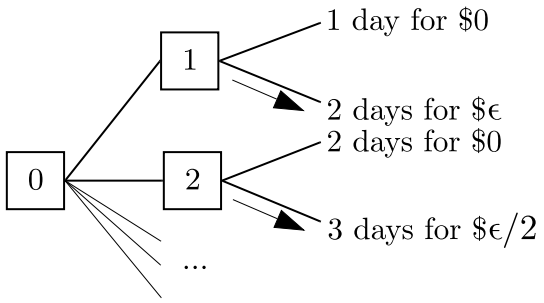


Figure 6-4. Pump against montonicity

The tree shows the number of days in Paradise and their prices. By a familiar argument, it follows, by pairwise guidance, that the agent should go down at all nodes following the initial one. And it follows from general nonemptiness and naïve choice that there is something the agent may do at the initial node. Hence, in two moves, a naïve agent would end up paying for something they could have had for free. This is a forcing money pump.

[45] Why not use monetary outcomes like before? When choosing from noncompact sets of monetary outcomes the agent automatically ends up paying for something free. For example, if from $A = \{\$1, \$2, \$3, ...\}$, they choose \$1, they pay \$1 relative to \$2. So, any such decision problem is automatically a money pump, in Cantwell's sense. The problem also arises if we understand the problem with pumps in terms of covering, since, arguably, all outcomes in A are covered.

Again, the trouble seems to be with noncompactness, and with general nonemptiness specifically, suggesting that the EverBetter Problems at issue are rational dilemmas. If we reject general nonemptiness, we cannot say that the agent is permitted to do something at the tree's initial node. *A fortiori*, we cannot say that they are permitted to make moves by which they end up paying for something free.

We also see that it is not monotonicity itself that is the problem. A naïve agent would presumably go to Paradise forever, if we made that outcome available in the tree above, thus making the set of available outcomes compact. They would thereby avoid getting pumped. Hence, money-pump arguments appear to overgenerate in the infinite noncompact case but not necessarily in the infinite compact case, where they can work to support transfinite transitivity and acyclicity.

7 Satan's Apple

At last, we get to Satan's Apple, a decision puzzle often used as a test case for theories dealing with infinite decision problems:[46]

> *Satan's Apple.* Eve prefers more apple to less but prefers to stay in Eden rather than have any apple outside of it. Satan can cut the apple as thinly as he wants. At 12:00 he offers Eve 50% of the apple. If she accepts, he offers Eve another 25% of the apple. And so on. If she refuses at any point, there are no further offers and she can keep as much of the apple as she got. If Eve accepts all the offers before 13:00, she is expelled from Eden for her greed.[47]

This superficially seems like a counterexample to transfinite acyclicity. It seems that the outcomes of Eve's actions are getting better and better yet in the limit they are worse. So, seemingly, we have a transfinite cycle of dispreference. Yet Eve's desires seem reasonable: she wants more apple, and she wants to be in Eden. Does this show that something is wrong with transfinite acyclicity after all? This section will argue that it does not, and instead use the distinctions developed in this paper to explain what's going on in Satan's Apple.

The first thing to note is that Satan's Apple is a supertask story. We therefore need to ask: "Do the outcomes of Eve's actions at each stage converge to that end-state in the relevant sense?"

While there is a good sense in which the actions of taking more and more slices have the action of taking the whole apple as their limit, it is not clear that the

---

[46] By Arntzenius et al. (2004) but also by Bartha et al. (2014) who introduce transfinite transitivity in their explanation of what's going on in Satan's Apple.

[47] This version is from Bartha et al. (2014).

outcomes of these actions have the *outcome* of taking the whole apple as their limit. That is, it is not clear that taking the whole apple *and being expelled from Eden* is the limit of taking more and more slices *while being in Eden*. After all, isn't there a discontinuity at the moment of expulsion?

This depends on how the story is implemented. One implementation is

> *Satan's Apple 2*. Imagine that Eve is magically removed from Eden at exactly 13:00 if she decided to accept all of Satan's offers beforehand.

Here the before-states do not seem to converge to the supposed end-state. For one thing, there is a physical discontinuity in Eve's position. Another implementation is

> *Satan's Apple 3*. Satan sets up a counter registering Eve's possession of the apple (0%, 50%, 75%, etc.) and a clock indicating how long Eve can stay in Eden after 13:00 (stay forever, leave by 13:30, leave by 13:15, etc.). At 12:00 Satan offers Eve 50% of the apple on the condition that the second clock goes from "forever" to "13:30". Since Eve only cares about the apple, given that she can eat it in Eden, she accepts and is moved a little bit closer to the exit. Then at 12:30 Satan offers her another 25% of the apple on the condition that the second clock goes from "13:30" to "13:15". Again, she accepts and is moved even closer to the exit. And so on.

In this version it is plausible that the before-states do converge to the supposed end-state (at least in physical terms), so at 13:00 she ends up with the apple but out of Eden.

But, in this case, it is not so hard to bite the bullet and say that Eve is unreasonable. She is willing to give up arbitrarily much of her Eden time for the sake of an arbitrarily small extra sliver of the apple, while she is also willing to

give up arbitrarily much of the apple for an arbitrarily short time back in Eden. It is Eve's stark trade-offs that land her in trouble in Satan's Apple 3. And it is not implausible to think they are to blame.

There remains the question of what Eve is to do in her situation. Satan's Apple is typically presented as a sequential decision problem similar to the infinite version of Tversky's pump considered above. As we saw there, rational agents are plausibly construed as *naïve* rather than *myopic*. Eve should decide what to do at any given time not just by looking at the trades immediately available to her, but rather by looking at all the possible outcomes she could obtain in her decision problem.

So, Eve should try to obtain the outcome that she considers choiceworthy from the set of all the possible outcomes. If the outcomes of taking more and more apple do *not* converge to the outcome of taking the whole apple (as in Satan's Apple 2), then Eve is arguably facing the problem of choosing from a noncompact set. And we saw that in noncompact option sets much of the standard theory of rational choice breaks down, so it is not implausible to think she is then facing a rational dilemma. If, on the other hand, the outcomes of taking more and more apple *do* converge to the outcome of taking the whole apple (as in Satan's Apple 3), then Eve has transfinitely cyclic preferences while choosing from a compact option set, and the fact that she ends up poorer in the end should bother her, but not us.

In their analysis of Satan's Apple, Bartha et al. (2014) suggest that Eve's story is a counterexample to money-pump arguments in general. In light of this section, we see that this is an overreaction. We should distinguish between agents with monotonic preferences in noncompact option sets (where noncompactness is to blame) and agents with transfinitely cyclic preferences in compact option sets

(where transfinite cyclicity is to blame). Satan's Apple is less of a problem than at first appears.

8 Conclusion

This paper introduced principles of transfinite transitivity and acyclicity, showed how to apply them in infinite-population ethics and variable-population ethics, and established their role for rational choice from infinite but compact option sets.

In that context the two transfinite principles can play the role that their finite namesakes play in finite option sets (with one major caveat regarding transfinite acyclicity): they help ensure the existence of choiceworthy options, are implied by consistency conditions on choice, and can be supported by means of money-pump arguments. They are also unthreatened by puzzles such as Satan's Apple.

This paper did not argue that we should believe in ordinary transitivity and acyclicity, nor that we should believe in them because of what they can do for us in the theory of rational choice. But it did argue that, if we do both, then we should also believe in transfinite transitivity and acyclicity for the same reasons.[48]

---

[48] There are other reasons to believe in transitivity. For example, Broome (2004) thinks that relations expressed by comparatives are transitive as a conceptual matter: "'A is more F than B' means that the degree to which A has the property F is greater than the degree to which B has this property, and the relation 'greater than' is transitive" (51). Perhaps transfinite transitivity can be supported in this way, too. Take the "higher than" relation, "higher than" being a comparative of "high". Say that x is higher than y iff x's height is greater than y's height, where heights are represented by points in $\mathbb{R}$. We also have a grasp on the notion of closeness in height, at least relative to a context. This can give us a notion of convergence. So, plausibly, the "higher than" relation maps to the $\leq$ relation on $\mathbb{R}$ not just with respect to order but also topology. And we saw in footnote 32 that $\leq$ on $\mathbb{R}$ is transfinitely transitive. Hence, the "higher than" relation is transfinitely transitive, too. Similarly for other relations expressed by comparatives.

9 References

Aumann, R. J. (1962). Utility theory without the completeness axiom. *Econometrica, 30*(3), 445-462. doi:10.2307/1909888

Allis, V., & Koetsier, T. (1995). On some paradoxes of the infinite II. *The British Journal for the Philosophy of Science, 46*(2), 235-247. doi:10.1093/bjps/46.2.235

Arntzenius, F., Elga, A., & Hawthorne, J. (2004). Bayesianism, infinite decisions, and binding. *Mind, 113*(450), 251-283. doi:10.1093/mind/113.450.251

Arrhenius, G. (2000). An impossibility theorem for welfarist axiologies. *Economics and Philosophy, 16*(2), 247-266. doi:10.1017/S0266267100000249

Bartha, P., Barker, J., & Hájek, A. (2014). Satan, saint peter and saint petersburg. *Synthese, 191*(4), 629-660. doi:10.1007/s11229-013-0379-9

Birchenhall, C. R. (1977). Conditions for the existence of maximal elements in compact sets. *Journal of Economic Theory, 16*(1), 111-115. doi:10.1016/0022-0531(77)90126-0

Blackorby, C., Bossert, W., & Donaldson, D. (2001). Population ethics and the existence of value functions. *Journal of Public Economics, 82*(2), 301-308. doi:10.1016/S0047-2727(00)00135-3

Bradley, B. (2006). Against satisficing consequentialism. *Utilitas, 18*(2), 97-108. doi:10.1017/S0953820806001877

Broome, J. (2003). Representing an ordering when the population varies. *Social Choice and Welfare, 20*(2), 243-246. doi:10.1007/s003550200175

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

Brown, C. (ms). *Better than nothing.* Unpublished manuscript. https://philpapers.org/rec/BROHTL

Buchak, L. (2013). *Risk and rationality.* Oxford: Oxford University Press.

Campbell, D. (1985). Impossibility theorems and infinite horizon planning. *Social Choice and Welfare, 2*(4), 283-293. doi:10.1007/BF00292691

Cantwell, J. (2003). On the foundations of pragmatic arguments. *Journal of Philosophy, 100*(8), 383-402. doi:10.5840/jphil2003100826

Carosi, L., & Zaffaroni, A. (1999). On the existence of maximal elements for partial preorders. *Journal of Information and Optimization Sciences, 20*(2), 271-286. doi:10.1080/02522667.1999.10699417

Davidson, D., Mckinsey, J. C. C., & Suppes, P. (1955). Outlines of a formal theory of value, I. *Philosophy of Science, 22*(2), 140-160. doi:10.1086/287412

Debreu, G. (1954). Representation of a preference ordering by a numerical function. In R. M. Thrall, C. H. Coombs & R. L. Davis (Eds.), *Decision processes* (pp. 159-166). New York: Wiley.

Devaney, R. L. (1989). *An introduction to chaotic dynamical systems* (2nd ed.). Redwood City, Calif; Wokingham: Addison-Wesley.

Diamond, P. (1965). The evaluation of infinite utility streams. *Econometrica, 33*(1), 170-177. doi:10.2307/1911893

Earman, J., & Norton, J. (1996). Infinite pains: The trouble with supertasks. In Adam Morton, & Stephen P. Stich (Eds.), *Benacerraf and his critics* (pp. 231-261) Blackwell.

Gillies, D. B. (1959). Solutions to general zero-sum games. In A. Tucker, & R. Luce (Eds.), *Contributions to the theory of games IV* (pp. 47-85). Princeton: Princeton University Press.

Grünbaum, A. (1970). Modern science and zeno's paradoxes of motion. In Wesley C. Salmon (Ed.), *Zeno's paradoxes* (pp. 200-250). Indianapolis/Cambridge: Bobbs-Merrill.

Gustafsson, J. E., & Espinoza, N. (2010). Conflicting reasons in the small-improvement argument. *Philosophical Quarterly, 60*(241), 754-763. doi:10.1111/j.1467-9213.2009.648.x

Hewitt, E. (1960). The role of compactness in analysis. *The American Mathematical Monthly, 67*(6), 499-516. doi:10.2307/2309166

Hurka, T. (1990). Two kinds of satisficing. *Philosophical Studies, 59*(1), 107-111. doi:10.1007/BF00368395

Kelley, J. L. (1955). *General topology*. Princeton, N.J; London: Van Nostrand.

Kreps, D. M. (2013). *Microeconomic foundations*. Princeton, NJ: Princeton University Press.

Kukushkin, N. S. (2008). Maximizing an interval order on compact subsets of its domain. *Mathematical Social Sciences, 56*(2), 195-206. doi:10.1016/j.mathsocsci.2008.01.003

Laraudogoitia, J. P. (2016). Supertasks. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy* (spring 2016 edition).

Lauwers, L. (1997). Continuity and equity with infinite horizons. *Social Choice and Welfare, 14*(2), 345-356. doi:10.1007/s003550050070

McClennen, E. F. (1990). *Rationality and dynamic choice: Foundational explorations*. Cambridge: Cambridge University Press.

Meacham, C. (2010). Binding and its consequences. *Philosophical Studies, 149*(1), 49-71. doi:10.1007/s11098-010-9539-7

Miller, N. R. (1980). A new solution set for tournaments and majority voting: Further graph- theoretical approaches to the theory of voting. *American Journal of Political Science, 24*(1), 68-96. doi:10.2307/2110925

Mukherji, A. (1977). The existence of choice functions. *Econometrica, 45*(4), 889-894. doi:10.2307/1912679

Parfit, D. (1984). *Reasons and persons* Oxford University Press.

Pollock, J. L. (1983). How do you maximize expectation value? *Noûs, 17*(3), 409-421. doi:10.2307/2215257

Pruss, A. R. (2018). *Infinity, causation, and paradox*. Oxford: Oxford University Press.

Sen, A. (1969). Quasi-transitivity, rational choice and collective decisions. *The Review of Economic Studies, 36*(3), 381-393. doi:10.2307/2296434

Sen, A. (1970). *Collective choice and social welfare*. San Francisco: Edinburgh: Holden-Day; Oliver & Boyd.

Sen, A. (1971). Choice functions and revealed preference. *The Review of Economic Studies, 38*(3), 307-317. doi:10.2307/2296384

Sen, A. (1993). Internal consistency of choice. *Econometrica, 61*(3), 495. doi:10.2307/2951715

Slote, M. (1989). *Beyond optimizing: A study of rational choice* Harvard University Press.

Smith, T. (1974). On the existence of most-preferred alternatives. *International Economic Review, 15*(1), 184-194.

Sorensen, R. (1994). Infinite decision theory. In J. Jordan (Ed.), *Gambling on god: Essays on pascal's wager* (pp. 139-159). Lanham, Md.: Rowman & Littlefield.

Thomson, J. F. (1954). Tasks and super-tasks. *Analysis, 15*(1), 1-13. doi:10.2307/3326643

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review, 76*(1), 31. doi:10.1037/h0026750

Uzawa, H. (1956). Note on preference and axioms of choice. *Annals of the Institute of Statistical Mathematics, 8*(1), 35-40. doi:10.1007/BF02863564

Van Liedekerke, L. (1995). Should utilitarians be cautious about an infinite future? *Australasian Journal of Philosophy, 73*(3), 405-407. doi:10.1080/00048409512346741

Weatherson, B. (ms). *Solving an infinite decision problem.* Unpublished manuscript. http://brian.weatherson.org/idt.pdf