Essays on Beneficence



Patrick Williamson

St Cross College
University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

14th June 2026

Acknowledgements

I owe a very great debt to my two wonderful supervisors, Andreas Mogensen and Alison Hills. Andreas and Alison both went far beyond the call of duty – doing so with great kindness, with inexhaustible patience, and with tremendous depth of insight.

Many others were generous with either conversation or written feedback, including Zach Barnett, Christopher Bottomley, John Broome, Krister Bykvist, Lucius Caviola, Mattia Cecchinato, Emma Curran, Lorenzo Elijah, Tomi Francis, Helen Frowe, Jonas Haeg, Riley Harris, Jonas Hertel, Joel Joseph, Ryan Kulesa, Kida Lin, Harry R. Lloyd, Kirsten Mann, Theo Naylor, Gary O'Brien, Theron Pummer, James Robinson, Pamela Robinson, Thomas Sinclair, Rhys Southan, H. Orri Stefánsson, Leora Sung, Teruji Thomas, Elliot Thornley, David Thorstad, Hayden Wilkinson, Lewis Williams, Timothy Luke Williamson, and Yoshiki Yoshimura.

I'm very thankful to have been a part of the Global Priorities Institute, a philosophical home which always struck me as something of a Platonic ideal. I'm extremely appreciative, too, of the chance to join St Cross College as one of their Uehiro Scholars for Future Generations in the academic year 2024/25.

Some of the terrific friends I found in Oxford (all booksmiths, of a sort) include Victoria Albacete, Camille Bernard, Greer Claybrook, Daphnée Guillemaud, Heidi Lane, Gemma Middleton, and Freya Norman. I still struggle to find the words, despite my best efforts, to express how grateful I am for the constant friendship of Jonathan Tjandra.

Perseverance was made possible by the loving kindness of my parents, Dugald and Rosemary Williamson, and by the encouragement of my philosophically wise brother, Timothy Luke Williamson.

This thesis is dedicated to my wife, Bethany Jedlicka, who fills each of my days with happiness, and who has supported me most of all.

Abstract

The essays in this collection examine our obligations of *beneficence*: that is, our obligations to improve the welfare of others, even when the 'others' in question turn out to be very distant strangers. I begin, in Chapter 1, by developing a novel account of our obligations of beneficence. I argue for the distinctive claim that each of us is required to start our adult moral life by committing towards an *altruistically-focussed lifeplan*. Such lifeplans are characterised, in the most general terms, by a radical and consuming devotion towards altruistic kinds of behaviour. I show how this result follows from an ecumenical set of premises. And I show how this result offers to illuminate various existing debates within the literature – most notably, the familiar debate over the extent to which we must each sacrifice our disposable income towards charity. In subsequent chapters I go on to make various concrete claims about how we should engage in beneficence. In Chapter 2, I explore a novel autonomy-based argument for thinking that it is generally permissible to costlessly rescue the lesser number. Although intuitions may differ, I am tempted to respond to the argument by rejecting widespread beliefs about the importance of moral autonomy. In Chapter 3, I argue that it is often permissible to display *altruistic near-bias*: that is, to display a special moral concern for the nearer-future welfare of others. I suggest that altruistic near-bias is generally permissible when distributing benefits across generations, but generally impermissible when distributing benefits across the life of a single beneficiary over time. In remaining chapters I pay increased attention to questions of risk and uncertainty. In Chapter 4 I argue that, once we render popular *partially aggregative* principles sensitive to considerations of risk and uncertainty, those principles may have an important role to play in mitigating against the demandingness of our all-things-considered moral theory. In Chapter 5, I argue in favour of *ex ante deontology*: I argue that, unless our decision is in the expected interests of all, we may not harm some as a method of benefitting others. Finally, in Chapter 6, I show how *ex ante* ethical standards such as this one can be saved from a recently influential puzzle concerning *ex ante* designation.

Word Count: 72,700

Table of Contents

Acknowledgements ii
Abstract iii
Table of Contents iv

0. Introduction **1**
0.1 Saints and Ostriches 1
0.2 Overview 2
0.3 Framing Assumptions 9

1. Beneficence as a Lifeplan Obligation 13
1.1 Introduction 13
1.2 The Lifeplan Argument 15
1.3 Premise (1) 17
1.4 Premise (2) 20
1.5 Premise (3) 29
1.6 Practical Upshots 32
1.7 Replies 35
1.8 Compositional Autonomy 42
1.9 Conclusion 51

2. Moral Autonomy and Lesser Numbers Permissions 55
2.1 Introduction 55
2.2 Moral Autonomy 58
2.3 Generating Lesser Numbers Permissions 64
2.4 The Similarity Objection 74
2.5 Existing Arguments 78
2.6 Conclusion 82

3. Two Kinds of Altruistic Near-Bias 88
3.1 Introduction 88
3.2 Against Intrapersonal Near-Bias 91
3.3 In Favour of Interpersonal Near-Bias 99
3.4 Objections 105
3.5 The Consistency Constraint 112

3.6 Conclusion 116

4. Partial Aggregation, Charitably Construed 119
4.1 Introduction 119
4.2 Partial Aggregation and Risk 122
4.3 *Ex Ante* Partial Aggregation and Charity 127
4.4 *Ex Post* Partial Aggregation and Charity 134
4.5 Objections 137
4.6 Partial Aggregation and Demandingness 140
4.7 A Puzzle and a Dilemma 142
4.8 Conclusion 150

5. Constraints, Consent, and Time 155
5.1 Introduction 155
5.2 The Diachronic Objection to *Ex Ante* Deontology 160
5.3 *Ex Post* or *Ex Ante* Deontology? 176
5.4 Conclusion 190

6. A Theory of *Ex Ante* Designation 193
6.1 Introduction 193
6.2 *Ex Ante* Ethics and the Problem of Multiple Designations 195
6.3 A Theory of *Ex Ante* Designation 199
6.4 The Informational Criterion 212
6.5 Criterial Conflict 219
6.6 Concluding Remarks 221

7. Concluding Remarks 225

0.
Introduction

1. Saints and Ostriches

Being a human is hard, I find. On the one hand I care about myself. Indeed, I care about myself very deeply. In my baser moments I might even go so far as to say: there is nothing for myself I wouldn't do.

On the other hand I care about other people. The list, here, is rather long. It includes the obvious suspects (my wife, for instance). But it also includes people I'll never even have the chance to meet: children suffering from communicable diseases on the other side of the globe, for instance, as well as future generations yet to be born for many hundreds of years to come.

What really makes things difficult, of course, is that my own interests are liable to conflict with the interests presented by others. This is true even if we put aside all potential conflicts between 'kith and kin.' For every dollar I ever spend on myself, I could have instead donated that dollar towards an effective international charity. And for every dollar I donate towards an effective international charity, I could have instead spent that dollar on myself. As I say, being a human isn't easy – it seems to involve constantly being torn between two competing worlds of concern.

How should we respond to this unfortunate predicament? An *ostrich* responds to the suffering of distant others by putting her head in the sand: that is, by simply ignoring the suffering of distant others altogether or pretending that it does not exist. It's quite easy to choose the path of the ostrich. It's quite hard, however, to choose the path of the ostrich without occasionally waking up to feelings of guilt and remorse. Surely the suffering of others *does* matter; surely we *are* required to do our small part in reducing it.

A *saint* responds to the suffering of distant others by emptying herself entirely: that is, by giving every second of her time and every cent of her savings to improve the plight of distant others. It's extremely difficult to choose the path of sainthood, even though some of our greatest moral heroes might come close to such a path. Intuitively, it seems that the path of sainthood is something *morally supererogatory*: to walk such a path is laudable but nevertheless beyond the call of one's moral duty.[1]

If we are morally required to become more than ostriches, but less than saints, then how much of our lives are we morally required to give away for the sake of distant others? Furthermore, when we do give away various parts of our lives for the sake of distant others, how should we go about doing so? Should we, for instance, show equal concern to other persons regardless of the time or place in which they happen to be born? May we, for instance, do harm to some as a means of benefitting others? As I'll label them, these are questions about the limits and nature of beneficence.

2. Overview

The following papers each progress our understanding of both the limits and nature of beneficence. Although the papers are independent enough to be read separately, they are nonetheless unified by some close thematic connections which will quickly become apparent. Broadly speaking, the papers begin by asking about the limits of beneficence: that is, by asking *to what extent* we must devote ourselves towards improving the welfare of others. The papers then turn to ask about the nature of beneficence: that is, by asking *how* we should devote ourselves towards improving the welfare of others. One important lesson to be learned, however, is that it may not always be possible to draw a clean line

[1] The distinction between saints and ostriches harkens to Kagan's (1989: 1–10) distinction between 'extremists' and 'minimalists.'

between these two types of questions. Sometimes, for instance, questions about the nature of beneficence will have direct bearing upon questions about the limits of beneficence. At the very least, the line between these two types of question will often turn out to be murkier than things first appear.

I begin, in Chapter 1, by setting out a novel account of our obligations of beneficence. I argue for the distinctive claim that each of us is morally required to begin our adult moral lives by committing ourselves towards an *altruistically-focussed lifeplan*, where such lifeplans are characterised, in general terms, by a radical and consuming devotion towards the promotion of welfarist benefits for other persons. In this way, I prove to be a closer friend of the saints than of the ostriches.

I argue that this conclusion follows from three independently plausible premises. The first premise is that making such lifeplan commitments promotes extreme expected good for others. The second premise is that, for a very large class of cases, making such lifeplan commitments is not expectedly worse for the one who does so. And the third premise is that, in general, when one is confronted with a lifeplan distinguished by these two features, one is morally required to make a serious and intentional commitment towards that lifeplan. These three premises, I argue, are enough to generate the relevant conclusion that each of us must begin our adult moral life by committing ourselves towards an altruistically-focussed lifeplan.

Having demonstrated my central conclusion I show that a similar conclusion holds even once we have qualified our discussion to take account of important non-consequentialist considerations – considerations such as partiality, integrity, psychological difficulty, and moral autonomy. Crucially, I also show how this central result sheds light on various existing debates about our obligations of beneficence. I pay particular attention, here, to the familiar debate over just how much of their disposable

income the affluent must donate towards charity. My key suggestion goes as follows: if I was required to begin my moral life by committing myself towards an altruistically-focussed lifeplan and previously failed to do so, my current obligations of charitable giving are likely to be far more strenuous than our intuitions first suggest.

As foreshadowed, I then turn to ask various concrete questions about the nature of beneficence. In Chapter 2, I investigate a novel argument in favour of *lesser numbers permissions*: that is, in favour of moral permissions to costlessly save the few rather than the many. I show, to be exact, that popular and commonsensical beliefs about the importance of 'moral autonomy' would in fact entail lesser numbers permissions within rescue cases such as:

> **Island.** Complete stranger Alice is dying on one island. Complete strangers Blaise, Charlie, Daniella, Ethan, and Francis are dying on a different island. You only have time to visit one such island. As a result, you can costlessly save some but not all.[2]

The notion of moral autonomy references, here, the popular sentiment that decision makers should be afforded substantial levels of creative freedom when it comes to deciding what sorts of distinctive valuable elements their future lives should contain.[3] Considerations of moral autonomy entail lesser numbers permissions in cases like **Island** – to give away the punchline for which I'll later argue – because it is very difficult to deny that saving the lesser number in such cases causes one's future life to contain 'distinctive valuable elements.'

There can be reasonable disagreement, I think, about how to interpret this central result. But my own sense is that, since it is implausible to maintain that there are lesser

[2] Anscombe (1967), Foot (1967), and Taurek (1977).
[3] With important articulations being Slote (1985), Shiffrin (1991), Kamm (1992), Hurley (1995), Lazar (2019), Pummer (2023; 2025), Tucker (n.d.).

number permissions in cases like **Island**, we should consider rejecting popular and commonsensical beliefs about the importance of moral autonomy.

In Chapter 3, I change course. I ask whether we must show equal concern to beneficiaries across time. In particular, I ask whether it is permissible to be *altruistically near-biased*: that is, to show special moral attention to the nearer-future welfare of beneficiaries. My key suggestion is that whether or not altruistic near-bias is permissible depends on the kind of choice context in which it is being exhibited. We generally lack moral reasons to be near-biased when distributing benefits across the life of a single beneficiary (the *intrapersonal case*). But we generally possess moral reasons to be near-biased when distributing benefits across the lives of separate and temporally subsequent persons (the *interpersonal case*).

The argument, here, contains several steps. The first step is to show that there are not good moral reasons for exhibiting altruistic *intrapersonal* near-bias. The second step is to show that there are, by contrast, good moral reasons for exhibiting altruistic *interpersonal* near-bias. The third step, roughly speaking, is to show that the reasons supporting one kind of near-bias cannot be extended in support of the other (and by extension that the criticisms against one kind of near-bias cannot be extended against the other). This leaves us, as I label it, with an 'asymmetric' approach towards the permissibility of altruistic near-bias.

I suggest that, insofar as we endorse an asymmetric approach to altruistic near-bias, this yields important implications for recent debates concerning our obligations of charitable giving. In particular, it suggests that whether or not a 'consistency constraint' governs charitable giving – a constraint that we donate on a regular basis rather than saving at interest in order to give greater quantities away later – depends in part on the types of charitable institutions to which one is giving.

In the second half of the thesis, I increasingly occupy myself with questions concerning risk and uncertainty. In Chapter 4, I turn to consider popular *partially aggregative principles*: principles according to which we should aggregate claims for aid, but only insofar as the claims in question are of sufficiently similar magnitudes. Such principles attempt to capture the following sorts of appealing verdicts: we ought to spare one person from death rather than a billion persons from slight headaches, even if we ought to spare (say) a billion persons from lost arms rather than one person from death.

Elsewhere, like others, I have expressed scepticism about the potential implications of partially aggregative principles.[4] In Chapter 4, however, my aim is more constructive. I aim to show how partially aggregative principles, when rendered suitably sensitive to questions of risk and uncertainty, may have an important role to play in protecting us from an *overly demanding* conception of morality: roughly speaking, a conception of morality according to which we must forfeit every marginal cent of our income and every marginal second of our time towards charity. I argue that partially aggregative principles often permit individuals to *not* make (even extremely effective) international aid donations. This is because those donations come at cost to the (identified) giver herself, and because such donations either (i) promote benefits for unidentified beneficiaries, or (ii) promote the good via the realisation of 'tipping points.' I show how, in each of these contexts, risk-sensitive versions of partial aggregation will generally generate a permission for givers to refrain from donating.

I stop short, however, of suggesting that partially aggregative principles could be seen as furnishing us with a complete solution to the problem of demandingness. This is in light of both a puzzle and a dilemma. The puzzle is that partially aggregative principles only appear to solve the overdemandingness problem in a manner which is extremely

[4] See my Williamson (forthcoming).

modally fragile, where we as yet lack the resources for evaluating whether a modally fragile solution to the problem of demandingness can be seen as an acceptable solution *simpliciter*. And the dilemma is that, depending upon the way in which partial aggregationists weight their reasons, the partial aggregationist must either choose between a highly demanding all-things-considered conception of morality or a particularly vicious version of the identified lives bias. Fully resolving these problems will, I suspect, require further work on underexplored elements of the demandingness problem and of the partially aggregative framework.

In Chapter 5, I consider whether we may harm some in order to benefit others. Famously, proponents of deontic constraints argue that the answer to this question is *no*: we may not harm some in order to benefit others, even when the benefits outweigh the harms by some considerable margin. The catch, unfortunately, is that it has proven notoriously difficult to specify how deontic constraints should operate in risky contexts.

My key concern, in Chapter 5, is to defend *ex ante* deontology: the view that we may only harm some for the sake of others insofar as doing so is in the expected interests of every single person. In particular, I respond to the recent objection that *ex ante* deontology cannot adequately handle choices made over time.[5] My response to this objection, in brief, is that certain diachronic versions of *ex ante* deontology do indeed deliver the right verdicts for the right reasons: sophisticated versions of *ex ante* deontology ensure that we never lethally and non-consensually impose ourselves upon the body of the one for the sake of the five. I conclude the discussion in Chapter 5 with a speculative proposal. Although recent arguments against *ex post* deontology have generally gone awry, we may indeed have promising theory-neutral grounds for preferring an *ex ante* deontological

[5] As in Kowalczyk (2022).

model over an *ex post* one: the intuitions underpinning the *ex post* model, I suggest, are more easily explained away.

Finally, in Chapter 6, I turn to do some foundational work concerning the ethics of risk and uncertainty. As will already be apparent from previous discussions, a popular approach to risky ethical decision making involves taking the *ex ante* perspective. According to *ex ante* ethical principles we ought to pay special attention, in some distinctive sense, to the *prospects* held by individual patients at the moment of choice. A patient's *prospects* under some act φ are to be identified with the probability weighted sum of their welfare under each of the possible outcomes which might follow φ's performance (allowing the probabilities in question to be transformed, perhaps, by some suitable function representing risk aversion).

One live challenge facing *ex ante* ethical principles is that there are different ways in which decision makers might designate patients at the moment of choice, where changing the mode of designation can change whether or not one's action qualifies as permissible under *ex ante* standards. This is the puzzle of *ex ante* designation.[6] In Chapter 6, I respond to this puzzle by developing a *normative theory of ex ante designation*: that is, a theory which specifies how decision makers ought to name patients when allocating prospects at the moment of choice. I argue that decision makers ought to use designators which are conducive to singular thoughts about beneficiaries. And I argue that decision makers ought to use designators carrying 'identity-relevant information.' I argue that each of these criteria are supported by and underlying normative rationale, and that, together, they routinely pick out a unique set of designators which ought to be used as the basis for

[6] Discussed at greatest length in Mahtani (2017; 2021).

the allocation of prospects. If this is right, then the puzzle of *ex ante* designation dissipates. The prospects for *ex ante* ethical principles end up looking quite good.

3. Framing Assumptions

I've provided a bird's-eye overview of the material to follow. Beyond this, I want to keep any lingering remarks about methodology fairly brief. Indeed, my methodology in what follows is an exceptionally predictable one: I attempt to demonstrate important conclusions about our obligations of beneficence on the basis of validly placed and well-justified premises. As I go, I generally aim to achieve a state of reflective equilibrium, a state in which any proposed moral principles are suitably "coherent with the moral convictions we share and have confidence in, including our convictions about general moral principles."[7]

It may be especially helpful, however, to draw attention towards some of the framing assumptions which underpin the following discussions. First, the following discussions investigate our obligations of *beneficence*: roughly speaking, our moral obligations to improve the lives of others with respect to considerations of welfare. The discussions focus especially upon our obligations to improve the welfare of some of those distant strangers who need help the most. A consequence of this framing is that I don't usually turn my attention towards different kinds of other-directed obligations we might hold – obligations of justice, for instance. The point of this framing assumption isn't to imply that such other-directed obligations don't exist, but only to make an impossibly large topic more tractable than it would have otherwise been.[8]

---

[7] Hooker (2000: 9), following Rawls (1980: 534). I am less insistent than some, such as Kagan (1989: 14), that such a process must be governed by a prohibition against 'danglers': that is, a prohibition against intuitive judgements which cannot be supported by justifications found one floor lower down. To defend these methodological sympathies would itself, of course, prove a book-length project.

[8] Of course, the relationship between welfare and justice is itself a difficult topic. For the view that welfarist considerations just are the currency of distributive justice, for instance, see Arneson (2000).

Second, in what follows, my focus is uniformly on our *individual* obligations of beneficence: that is, on the obligations that we as individuals hold to improve the lot of others with respect to considerations of welfare. My focus generally isn't on the obligations of beneficence which might be held by (say) corporations or government bodies or community collectives, if such things can be said to hold obligations of beneficence. Again, this assumption only exists for the sake of tractability. I don't mean to imply, in making it, that things such as institutional obligations of beneficence couldn't exist.[9]

Third, for a while at least, I will march under the banner of ecumenicalism. I will demonstrate claims about the limits and nature of beneficence while remaining neutral about which fundamental moral principles are correct at the level of bedrock. The conclusions I draw about beneficence in Chapter 1, for instance, follow from premises which should appeal to both consequentialists and non-consequentialists. Adopting ecumenicalism gives my conclusions an additional level of dialectical force: the relevant conclusions cannot be resisted, it seems, without giving up some deeply appealing and widely-held moral beliefs.[10]

Fourth, throughout, there is another important way in which I can remain ecumenical. Although my concern is with beneficence, which I take to be a matter of promoting welfare, I won't usually need to take a stance on which theory of welfare is most plausible. I won't usually need to take a stance, that is, on whether lives are made better by the presence of hedonistic pleasure, by the satisfaction of suitably pruned preferences, by the attainment of certain objective goods, or by some other important feature. This is because

[9] For seminal discussions on group agency and group moral responsibility, see French (1984), May (1987), Pogge (2002), Miller (2004), Pettit (2007), List & Pettit (2011), Tuomela (2012), and Dietz (2016).
[10] In this way, I am highly sympathetic to the argumentative style popularised by Singer (1972). Of course, as will quickly become apparent, the argument of Chapter 1 differs from Singer's in various important respects.

many of the key conclusions I wish to demonstrate are likely to hold regardless of whichever theory of welfare turns out to be correct. Chapter 1, for instance, draws on the premise that the giving of one's time and resources towards effective charities promotes the expected welfare of other persons to an extremely great degree. For the reasons I sketch, a premise like this holds under virtually all plausible candidate theories of welfare. Of course, where one's underlying theory of welfare does indeed bear on the plausibility of the conclusions I draw, I make sure to indicate as much within the text itself.

References

Anscombe, E. (1967) 'Who is Wronged? Philippa Foot and 'Double Effect'', *The Oxford Review, 5:* 16–17.

Arneson, R. (2000) 'Welfare Should Be the Currency of Justice', *Canadian Journal of Philosophy*, 30(4): 497–524.

Dietz, A. (2016) 'What We Together Ought to Do', *Ethics*, 126(4): 955–982.

Foot, P. (1967) 'The Problem of Abortion and the Doctrine of Double Effect', *Oxford Review*, 5: 5–15.

French, P. (1984) *Collective and Corporate Responsibility*. Columbia University Press.

Hooker, B. (2000) *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*. Oxford University Press.

Hurley, P. (1995) 'Getting Our Options Clear: A Closer Look at Agent-Centred Options', *Philosophical Studies*, 78(2): 163–188.

Kagan, S. (1989) *The Limits of Morality*. Oxford University Press.

Kamm, F. (1992) 'Non-Consequentialism, the Person as an End-in-Itself, and the Significance of Status', *Philosophy & Public Affairs*, 21(4): 354–389.

Kowalczyk, K. (2022) 'People in Suitcases', *Journal of Moral Philosophy*, 20(1–2): 3–30.

Lazar, S. (2019) 'Moral Status and Agent-Centred Options', *Utilitas*, 31(1): 83–105.

List, C. & Pettit, P. (2011) *Group Agency: The Possibility, Design and Status of Group Agents*. Oxford University Press.

Mahtani, A. (2017) 'The Ex Ante Pareto Principle', *The Journal of Philosophy*, 114(6): 303–323.

——— (2021) 'Frege's Puzzle and the Ex Ante Pareto Principle', *Philosophical Studies*, 178(6): 2077–2100.

May, L. (1987) *The Morality of Groups: Collective Responsibility, Group-Based Harm and Corporate Rights*. University of Notre Dame Press.

Miller, D. (2004) 'Holding Nations Accountable', *Ethics*, 114(2): 240–268.

Pettit, P. (2007) 'Responsibility Incorporated', *Ethics*, 117(2): 171–177.

Pogge, T. (2002) *World Poverty and Human Rights*. Polity Press.

Pummer, T. (2023) *The Rules of Rescue: Cost, Distance, and Effective Altruism*. Oxford University Press.

——— (2025) 'Compensated Altruism and Moral Autonomy', *Social Philosophy and Policy*, 42(1): 186–203.

Rawls, J. (1980) 'Kantian Constructivism in Moral Theory', *Journal of Philosophy*, 77(9): 515–572.

Shiffrin, S. (1991) 'Moral Autonomy and Agent-Centred Options', *Analysis*, 51(4): 244–254.

Singer, P. (1972) 'Famine, Affluence, and Morality', *Philosophy & Public Affairs*, 1(3): 229–243.

Slote, M. (1985) *Common-Sense Morality and Consequentialism*. Routledge.

Taurek, J. (1977) 'Should the Numbers Count?', *Philosophy & Public Affairs*, 6(4): 293–316.

Tucker, C. (n.d.) 'Autonomy: How Lesser Goods Can Be Justifying Reasons', unpublished manuscript.

Tuomela, R. (2012) 'Group Reasons', *Philosophical Issues*, 22: 402–418.

Williamson, P. (forthcoming) 'Relevant Claims and Causal Ramifications', *Australasian Journal of Philosophy*, 1–23.

1.

Beneficence as a Lifeplan Obligation

Abstract. Starting your adult moral life by committing towards an altruistically-focussed lifeplan is much better for other persons and it is no worse for you. Thus, you are morally required to start your adult moral life by committing towards an altruistically-focussed lifeplan. This is the core of the *Lifeplan Argument*, which I propose and defend in the following pages. In the latter half of the discussion I sketch some important practical implications which might follow from the Lifeplan Argument if sound. I also respond to potential objections. I pay particular attention to the concern that the Lifeplan Argument is incompatible with a robust concern for important non-consequentialist considerations such as partiality, integrity, psychological difficulty, and moral autonomy. I suggest that this concern is not well-founded. If this is right, we must each begin our moral lives by committing to become altruistically-focussed.

1. Introduction

The path of moral sainthood sure looks hard. There's blood, sweat, and tears. None of these three liquids ever seems to end. The path of moral ostrichhood looks much easier: simply putting our heads in the sand and pretending that the suffering of distant others doesn't exist.

In this article I don't argue that we are morally required to be moral saints. But I nonetheless argue for an important and revisionist claim in the vicinity. I argue that each of us is required to begin our adult moral lives by making a serious and intentional commitment towards the living out of an *altruistically-focussed lifeplan*: that is, a lifeplan

devoted in a radical and consuming way towards the pursuit of welfarist benefits for other persons. I argue for this central conclusion by developing a novel argument, which I label the *Lifeplan Argument*.

One might worry that, in developing the Lifeplan Argument, I'm attempting to prove a conclusion of only esoteric interest. After all, the vast majority of us are *not* at the start of our adult moral lives. We are somewhere in the middle of our adult moral lives, having already embarked upon those careers, projects, and attachments which we take to imbue our lives with meaning. Why should agents such as these care about what they were morally required to do (say) twenty or thirty years ago?

This way of thinking, I'll suggest, is deeply misguided. There are different ways in which a past failure to act upon one's obligations of beneficence might bear upon one's current obligations of beneficence. One obvious possibility is that agents who previously failed to act upon their obligations of beneficence might be liable to make reparations for their past failures, even when those reparations appear highly burdensome.[1] Another possibility is that agents might be required to modify or reshape those projects which came about as a causal consequence of their past failures. In these various ways, the Lifeplan Argument offers to shed light upon the obligations of beneficence held by agents across their lives as a whole.

The discussion is structured as follows. In §2, I set out the core structure of the Lifeplan Argument. In §3–5, I successively defend each of the Lifeplan Argument's key premises – premises which together generate the conclusion that each of us is required to begin our adult moral lives by committing ourselves towards an altruistically-focussed lifeplan. And in §6, I set out some practical implications which might follow from the Lifeplan Argument if sound.

[1] A possibility explored at greatest length in Barry & Øverland (2016: 30–49) and Cordelli (2018).

I conclude, in §7–§8, by responding to potential objections. My general concern, here, is to show that we cannot resist the Lifeplan Argument by pointing towards familiar and important non-consequentialist considerations – considerations such as partiality, integrity, psychological difficulty, and moral autonomy. I show, in particular, that a qualified version of the Lifeplan Argument is expressly compatible with a robust concern for considerations of moral autonomy. If this is right, we must each begin our adult moral lives by committing to become altruistically-focussed. Our ordinary conception of beneficence is in drastic need of revision.

2. The Lifeplan Argument

Without further ado, here is the core structure of the Lifeplan Argument:

> *The Lifeplan Argument*
> (1) At the start of one's adult moral life, committing to an altruistically-focussed lifeplan promotes extreme expected good for others.
> (2) At the start of one's adult moral life, committing to an altruistically-focussed lifeplan is not expectedly worse for the one who does so.
> (3) If committing to a lifeplan promotes extreme expected good for others without being expectedly worse for the one who does so, one is morally required to commit to that lifeplan.
> (4) Therefore, at the start of one's adult moral life, one is required to commit to an altruistically-focussed lifeplan.

The following several sections explain and defend premises (1)–(3), which together generate the conclusion that each of us is morally required to begin our adult moral lives by committing ourselves towards an altruistically-focussed lifeplan. First, though, some comments made at the bird's-eye level may be helpful.

Premise (1) of the argument claims that committing towards an altruistically-focussed lifeplan promotes extreme expected good for others. For reasons I sketch shortly, premise (1) is especially plausible. Nonetheless, premise (1) introduces some contested terms. In

particular, it introduces the notion of an 'altruistically-focussed lifeplan' and the notion of the 'start' of an adult moral life. Defending premise (1) will thus provide me with a chance to clarify my usage of these contested terms.

Premise (2) claims that committing to an altruistically-focussed lifeplan is not prudentially worse for the one who does so. Premise (2), unlike premise (1), is likely to strike many readers as highly controversial, if not trivially false. Surely altruistically-focussed lifeplans *are* worse for those who commit to them, and surely those lifeplans are worse precisely in virtue of their altruistically-focussed character.[2] In my view, however, there is an extremely strong case for thinking that it is not prudentially worse for agents to begin their moral lives by committing themselves towards altruistically-focussed lifeplans, at least for very many such lifeplans. I develop four independent arguments in favour of this claim, drawing upon both philosophical and empirical evidence in support of my case.

This leaves premise (3): the more general premise that one is required to commit towards those lifeplans which are much (expectedly) better for other persons and which are not (expectedly) worse for oneself. Whether or not we endorse premise (3) depends, of course, upon our underlying view about how exactly duties of beneficence are generated in the first place. My general strategy will be to explain why premise (3) is both modest and intuitively appealing, before then sketching a popular framework of moral reasons which would indeed lend support to claims in the vicinity of (3).

As we will see, there are similarities between the above Lifeplan Argument and certain other influential arguments in the literature for strenuous obligations of beneficence. I

[2] Much of the historical literature on demandingness presupposes that the typical characteristic features of altruistically-focussed lives can prove highly prudentially costly, including *inter alia* Singer (1972), Slote (1985), Kagan (1989), Unger (1996), Hooker (2000), Murphy (2000), Mulgan (2001), Ashford (2003), Cullity (2003; 2004), Noggle (2009), Timmerman (2015), Woollard (2015), Pummer (2016; 2023), and Thomson (2022).

particularly have in mind, here, historically influential arguments defended by Peter Singer and Peter Unger.[3] But while there are similarities, there are also key differences. As may already be clear, one key difference is that the Lifeplan Argument argues for a very distinctive *kind* of moral requirement – a requirement that we each begin our lives by making certain kinds of altruistic pre-commitments. In doing so the Lifeplan Argument comes with a distinctive set of practical implications. It also has the ability to elegantly escape a historically influential set of objections.

3. Premise (1)

My first claim is that committing to an altruistically-focussed lifeplan, at the start of one's adult moral life, promotes an extreme quantity of expected good for others. An initial challenge, here, lies in explaining what I mean in using these terms.

What do I mean by 'at the beginning of one's adult moral life'? This phrase is clearly a term of art. For current purposes, what matters is that moral agents extended through time typically have the ability somewhere in the earlier stages of their existence – early adulthood, say – where they can make careful and systematic plans about what sort of careers they will pursue, what sort of long-term personal projects they will develop, what sort of talents they will refine, what sort of relationships they will seek out, and so on. My claim is that, when agents find themselves at such a stage of life, committing to an altruistically-focussed lifeplan promotes extreme quantities of expected good for other persons.

Of course, from the agent's perspective, there might exist a substantial window of time, vague at the boundaries, which could be properly characterised as the 'start of adult moral

[3] Singer (1972) and Unger (1996).

life.' There might be many points within that window at which an altruistically-focussed lifeplan could be formed and adopted. This sensible point need not derail the current discussion. Premise (1) would then read that making a commitment towards an altruistically-focussed lifeplan at some moment within the relevant window promotes extreme quantities of expected good for other persons.

Accepting that there can be such a point (window) of time should not be taken to imply that persons prior to such a point, such as young children, exist in a nebulous vacuum entirely detached from partial attachments, commitments, personal projects, and so on. That would be a fairly absurd direction in which to move the discussion. The point is simply that, given the way in which human social life is generally structured, persons at the start of their adult moral lives often do find themselves with substantial levels of creative freedom in determining what sort of shape their long-term future will have.[4]

What, then, do I mean by an 'altruistically-focussed lifeplan'? This is, in short, a systematic and intentional plan to make one's future life unfold in a manner which is *altruistically-focussed*.[5] Of course, this only kicks the can further down the road. We are then in want of an explanation of what it means for one to live an altruistically-focussed life. Garrett Cullity initially described such lives as having the following distinctive character:

> I have an altruistically-focused life if a guiding aim of mine is to constrict my pursuit of my own fulfilment as much as I bearably can, for the purpose of benefiting others. Lives that are not like this—those that have an overall non-altruistic focus, and those that lack a single overall focus (as most lives do)—we can call "non-altruistically-focused."[6]

[4] Of course, young adults may have lacked this liberty within the vast majority of historical social arrangements, and still lack it today within certain cultural milieus. For sweeping discussion, see Taylor (1989; 1991).
[5] To the best of my knowledge, the helpful phrase 'altruistically-focussed' originates in Cullity (2003: 407).
[6] *Ibid.*

This is getting close to what I have in mind by the notion of an 'altruistically-focussed life.' But I do not think it is necessary to say that the guiding aim of an altruistically-focussed life should be to *constrict* one's own personal fulfilment as much as is bearable for the sake of others. There is nothing, I think, in our ordinary sense of an 'altruistically-focussed person' that demands pure asceticism. I think we do better to simply say that altruistically-focussed lives are characterised by being devoted, in a radical and consuming way, towards the promotion of benefits for other persons.

One might still want a more concrete definition of the 'altruistically-focussed life' than this. In particular, one might worry that characterising altruistically-focussed lives as being devoted in a radical or consuming way towards the promotion of benefits for other persons comes a little too close to giving mere synonyms for the phrase 'altruistically-focussed.' One potential solution, here, would involve setting a crude and imperfect benchmark which helps us filter out altruistically-focussed lifeplans from non-altruistically-focussed ones: perhaps, in order to qualify as an altruistically-focussed lifeplan, 80% of the professional and personal projects contained within a lifeplan must explicitly exist for the sake of bringing high-magnitude welfarist benefits to other persons (preventing deaths, severe illnesses, and so on). Another possibility, though, involves pointing out that we can often grasp the concept of an 'altruistically-focussed lifeplan' by ostension. Many of us are lucky enough to encounter persons who might properly be called moral saints: persons who shape their careers, their personal interests, and their interpersonal relationships towards the mitigation of the great moral horrors of our time, moral horrors such as global poverty. Those persons are living what I take to be canonical examples of altruistically-focussed lives. Such lives are indeed compatible with the presence of certain personal pleasures detached from the activity of altruism, but altruistic behaviour nonetheless qualifies as a central and defining focus of such lives.

With this definitional apparatus in place, I take it that premise (1) should qualify as fairly uncontroversial. It should be fairly uncontroversial that committing towards an altruistically-focussed lifeplan promotes extreme levels of expected good for others. Just to dwell upon one well-discussed example: under the best estimates of independent charity evaluators, each marginal $5500 donated to the Against Malaria Foundation (AMF) results in an additional life saved in expectation.[7] One does much more than raise $5500 for AMF when one pursues an entire career of advocacy and activism in favour of AMF. One does what I have called an extreme amount of expected good for others.[8]

4. Premise (2)

Premise (2) of the argument is much more controversial. Premise (2) claims, recall, that it is not expectedly *prudentially* worse for the agent when she starts her moral life by committing herself towards an altruistically-focussed lifeplan. Many philosophers will be tempted not merely to deny premise (2), but rather, to assert the exact opposite: to assert that altruistically-focussed lifeplans are worse for those who commit to them, and that these lifeplans are worse precisely in virtue of their altruistically-focussed character.[9]

Now, I grant that *certain* types of altruistically-focussed lifestyles seem prudentially bad in expectation – providing field aid to combatants within extremely brutal theatres of war, say. But we must not therefore generalise that *all* altruistically-focussed lives are

[7] For the various complexities, see https://www.givewell.org/charities/amf.
[8] There may be some altruistic organisations to which one could devote a career to without promoting 'extreme amounts of expected good' in the sense of (say) saving several lives in expectation or sparing several persons from sickness, disease, poverty, or severe mistreatment (in expectation). I do not take such organisations to be the focus of my discussion. Importantly, premise (1) is framed in expectational terms. Plausibly, various kinds of altruistic activity only effect good by contributing to the reaching of 'tipping points' (as in discussions of animal welfare activism in Singer (1980), Norcross (2004), and Kagan (2011)). An altruistically-focussed life devoted to such causes might carry a large chance of doing nothing but a small chance of doing enormous amounts of good. This complication shouldn't matter for current purposes, since I take it that we should promote the good in a broadly evidence-relative and stakes-sensitive manner. For related discussion, see Cullity (2004: 58–69).
[9] As in Wolf (1982).

prudentially worse for those who seek them out. For the Lifeplan Argument to succeed, all we need is that there exist *some* altruistically-focussed lifeplans which are not prudentially worse for those who seek them out. My point, in what follows, is that this criterion is almost certainly met by a surprisingly wide range of altruistically-focussed lifeplans.

It will help, in arguing the case, to have a concrete example as our focal point. Thus consider:

> **Frost's Choice**. You are at the start of your adult moral life. Just as the poem says, you face two diverging paths. You may choose one of the following.
>
> *Option 1*. Your friend notices your latent sketching talents and offers you a cosy job in her architectural firm, specialising in office buildings. You'd get a lot of enjoyment from various aspects of the work, such as the pursuit of aesthetic goals and the regular chance to engage in creative problem solving. The salary would be very comfortable. You would work very good hours, leaving you ample time to pursue your hobbies – gold-panning and basket weaving prime amongst them.
>
> *Option 2*. You could instead prepare to become a full-time medical worker just south of the Sahara, providing malarial vaccines to infant children who would have otherwise been utterly neglected. Your basic needs will usually be met. However, your living conditions will never be especially luxurious. Furthermore, you will find yourself unable to pursue as many long-term social connections as you would otherwise like, finding that such relationships sometimes detract time and attention from your work in a manner which leaves you riddled with guilt.

Frost's Choice is, of course, a highly idealised case. For now, simply note that *Option 2* is a canonical example of what I have labelled an 'altruistically-focussed lifeplan': it is devoted in a radical and rather consuming way to the promotion of welfarist benefits for other persons (preventing deaths, severe illnesses, and so on). *Option 1*, by contrast, given the above definitional apparatus, plausibly fails to qualify as altruistically-focussed.

It is tempting to assume from the armchair that it would be prudentially worse for you to commit to an altruistically-focussed lifeplan like *Option 2*. But on reflection, there are at least four independent arguments for thinking that lifeplans such as *Option 2* aren't

generally prudentially worse for those who seek them out – or at least, aren't prudentially worse for those who to commit to them at the start of their moral lives.

4.1 Parity Goods

Here is a warm-up argument. A choice between highly contrastive lifeplans like *Option 1* and *Option 2* is a classic example of a choice between candidate sets of *parity goods*. When we say that two objects *A* and *B* are evaluatively on a par, we mean to convey the following. *A* and *B* are in some broad sense comparable. But neither *A* nor *B* are better (worse) than the other. Neither are *A* nor *B* strictly equally as good as each other. Thus, as Ruth Chang puts it, the artistic talent of Mozart is comparable to that of Michelangelo, even though neither Mozart nor Michelangelo are straightforwardly better nor worse nor strictly equally good when held in direct comparison.[10]

Insofar as one entertains the possibility of parity, the comparison between *Option 1* and *Option 2* is just the sort of case in which we should expect a parity relation to emerge. *Option 1* brings a life filled with the distinctive goods of comfort, professional success, and aesthetic craftwork. *Option 2*, by contrast, unlocks a radically different life filled with the sense of meaning and purpose generated by the direct saving of lives. These are prudential goods which are difficult to compare in just the same way that the aesthetic goods produced by Mozart and Michelangelo were difficult to compare. If these bundles of goods are best characterised as standing in a parity relation, neither qualifies as strictly better nor worse than the other, and we arrive at premise (2). It would not be *worse* for you to pursue *Option 2*.

[10] Chang (2002). Parfit (1984: 431–432) and Griffin (1986: 96–98) previously refer to the relation of "rough equality."

Importantly, this line of argument can still have traction even if one rejects the exact conception of parity articulated by Chang. If, for instance, one thinks that our intuitions about parity in fact gain traction because certain value-theoretic comparisons are irreducibly vague, and if it follows from this that there are regularly truth-value gaps in comparative value claims of the form "*x* is *F*-er than *y*", then it can nonetheless follow that it is *not true* that *Option 2* is worse than *Option 1*.[11] One could then run a modified version of the Lifeplan Argument according to which one must pursue those altruistically-focussed lifeplans which are expectedly better for others and for which it is *not true* that those lifeplans are prudentially worse for the one who pursues them.

Now, the parity argument may have special appeal to those who endorse objective list theories of welfare. After all, the parity argument seems to draw upon the idea that there are irreducibly different *kinds* of goods which contribute to one's well-being, where these goods can compete with one another in a manner which is difficult to precisely evaluate. (If you feel that the prudential goodness of learning facts about the galactic structure of the Milky Way is difficult to compare to the prudential goodness of licking a tangy sherbet, then you are a prime target audience for the parity argument.) But it is worth emphasising that the parity argument doesn't presuppose belief in an objective list theory. The hedonist, for instance, may believe that a mental experience is good for me just to the extent that the experience in question strikes me as enjoyable, where the extent to which a mental experience strikes me as enjoyable is typically a function of multiple qualitative features including duration, intensity, and Millian 'nobility.'[12] Such a hedonistic picture is in principle compatible with the judgement that *Option 1* and *Option 2* are on a par. The overall enjoyable experience contained within *Option 1* may be longer but less 'noble' than

[11] See recent discussion in Broome (2022).
[12] Mill (1998). See also Crisp (2006).

the overall enjoyable experience contained within *Option 2*, such that no precise comparison can be made between the two bundles of experience.

4.2 Relevant Empirical Results

The above argument is highly philosophical. Without empirical evidence, it is likely to ring a little hollow. Let's turn in an empirical direction, then. I want to draw attention here to two empirical results which lend support to the claim that altruistically-focussed lifestyles like *Option 2* are not worse than alternatives such as *Option 1*.

First: although altruistically-focussed lifestyles like *Option 2* often seem to contain lesser financial means, it is very hard to argue from this result that altruistically-focussed lifestyles like *Option 2* must be *worse* for those who seek them out. Why? Well, a sweeping empirical literature has long examined the correlation between income and life-satisfaction. One lesson to be learned from this literature is that, although there is indeed a correlation between income and reported life satisfaction, this correlation is not *nearly* as marked as one might expect: in particular, the correlation between these two features is plausibly non-linear such that substantial decreases to the income of the affluent are only correlated with trace decreases in levels of reported life satisfaction.[13] This makes it much harder, in general, to argue that one's life will go substantially worse in *Option 2* in virtue of that life having lesser financial means. Of course, your reported life satisfaction might turn out to be much lower insofar as a given altruistically-focussed lifeplan required you to enter conditions of (or near) material poverty. But it is not obvious that a requirement to pursue *Option 2* or similar altruistically-focussed lifeplans would, in general, require that the agent send herself into (or near to) conditions of material

[13] See, for instance, the logarithmic model presented in Stevenson and Wolfers (2013). For a survey of some of the relevant empirical work, see MacAskill, Mogensen, & Ord (2018: 185–189).

poverty: as is a familiar point, in order to efficiently and sustainably do good, altruists must take decent minimal care of themselves.

Second: if on the on hand there are surprisingly weak empirical correlations between income and life satisfaction, there are, on the other hand, some surprisingly strong empirical correlations between prosocial behaviour and life satisfaction – that is, between life satisfaction and the sorts of behaviours inherently characteristic of altruistically-focussed lives like *Option 2*. We are spoiled for choice, here, when it comes to picking examples. The regular giving of money or financial resources is universally associated with greater giver happiness: the word 'universal' indicates that the correlation has been found to hold in more than one hundred economically and culturally diverse countries.[14] Furthermore, there is strong evidence that this relation is causal, such that it is giving *itself* which makes givers happier.[15] A large number of studies have also examined the impact of formal (non-familial) volunteering behaviour upon quality of life, finding that formal volunteering correlates with markedly increased subjective reporting of health and happiness.[16] Again, there is strong evidence that the relation is causal, at least with respect to the impact of volunteering upon happiness.[17]

This suggests the following. If we want to resist premise (2), we will be wedged between a rock and a hard place. On the other hand we must try to resist an empirical literature indicating that income – at least across minimally comfortable lifestyles – yields surprisingly weak correlations with life satisfaction. On the other hand we must try to resist a large empirical literature demonstrating the expansive welfarist benefits

[14] Aknin et al. (2013). For a recent meta-analysis on the link between prosociality and well-being, see Hui et al (2020).
[15] *Ibid.*
[16] Borgonovi (2008).
[17] *Ibid.*

associated with altruistically-focussed activities – that is with, those activities inherently characteristic of lives like *Option 2*.

4.3 Testimony of the Saints

One way of figuring out whether an altruistically-focussed lifeplan like *Option 2* would be worse for you is to go and consult the testimony of those various 'saintly' persons who do in fact select lifeplans such as *Option 2*. It turns out that it is extremely common for 'saintly' persons to profess sentiments in the following direction: pursuing the life of sainthood did not turn out *worse* for them, it did not cause them to feel *regret*, and it was not something which they came to understand as a *loss*. What's especially surprising is the chirpy and chipper ways in which such testimony is regularly conveyed.[18]

This idea may sound surprising, but much of this surprisingness dissipates when we turn our attention towards particular accounts. An example helpfully foregrounded by Vanessa Carbonell is Tracy Kidder's detailed account of the life of Paul Farmer, one of the founders of Partners In Health (PIH), often cited as a paragon example of a modern moral saint. Speaking of her early encounters with Farmer, Kidder writes:

> … A few weeks later, I took [Farmer] to dinner … He had described himself as "a poor people's doctor," but he didn't quite fit my preconception of such a person. He clearly liked the fancy restaurant, the heavy cloth napkins, the good bottle of wine. What struck me that evening was how happy he seemed with his life. Obviously, a young man with his advantages could have been doing good works as a doctor while commuting between Boston and a pleasant suburb … The way he talked, it seemed he actually enjoyed living among Haitian peasant farmers. At one point, speaking about medicine, he said, "I don't know why everybody isn't excited by it."[19]

[18] For cases and interpretive discussion, see Carbonell (2009; 2012). C.f. Wolf (1982).
[19] Kidder (2003: 10).

As things seem to me, what is involved in taking such testimonies seriously is assigning a much higher level of credence to the notion that lifeplans such as *Option 2* are not typically expectedly worse for those persons who end up seeking them out.

At this point, alarm bells might be ringing. First, there could be a self-selection effect such that it is only the 'saintlike' – those who are naturally disposed to select *Option 2* – for whom the choice of *Option 2* actually makes their life go well. For the rest of us who lack naturally saintly dispositions, perhaps *Option 2* would indeed make our life go worse. Second, there are clearly strong instrumental reasons for the saints to not complain about their lot in life: insofar as you care about a particular altruistic cause, it is hardly a means of attracting interest in your cause to go around complaining about how burdensome it is.

Nonetheless, blanket scepticism strikes me as far too hasty. This is for the key reason that moral saints do not merely *assert* that their lives go well; rather, in making such claims, moral saints point to general mechanisms which give a plausible explanation of why their lives go well, where the mechanisms in question do not seem to gain traction by exploiting psychological features peculiar to the 'saintly.'[20]

Some of these mechanisms are the very sorts of mechanisms cited by researchers in attempting to give an explanation why prosocial behaviours such as volunteering tends to have a positive causal influence on happiness levels. One common suggested mechanism is that directly having the chance to eliminate severe suffering is a qualitatively satisfying experience.[21] Another is that regular prosocial behaviour of the kind which (say) Paul Farmer exhibited might embed one within an unusually strong set of social and

[20] See relevant discussion in Carbonell (2009).
[21] This, I take it, is the sentiment behind the quote popularly attributed to Farmer that he felt most alive when helping others (as reported in Kidder (2003: 188)).

professional relationships, where such relationships tend to make one's life flourish.[22] Each of these mechanisms are 'general' in that they do not work by exploiting features peculiar to the psychologies of moral saints. The presence of possible general mechanisms such as these suggests, I think, that we should not rush to reject saintly testimony out of hand, even if it is initially tempting to do so.

4.4 The Argument from Trying

Of course, inspired by saints like Farmer, you might set out upon a given altruistically-focussed lifeplan like *Option 2* only to quickly realise it is not for you. Perhaps you quickly realise that you have an unshakeable phobia to certain aspects of that life – injections, say. We should grant that, in this particular case, it could indeed be very bad for you to be 'locked in' to such a lifeplan for perpetuity.

Note, however, that the sort of commitment invoked by the Lifeplan Argument need not entail *perpetual locking in* of such a sort. I take it that the sort of commitment invoked by the Lifeplan Argument is to make a serious and concerted effort to make the rest of one's life unfold in a manner which is altruistically-focussed. Such a commitment would, presumably, be compatible with a permission to leave the life of doctoring if you eventually came to realise (several years *post facto*, say) that such a life contained features with which you could not be reconciled. In light of this fact it becomes much harder to argue that the commitment in question is expectedly worse *for you.*[23]

One might think that I am underselling, here, the prudential costs involved in (say) trying out *Option 2* for a stretch of several years. Couldn't this choice cost you a great

[22] Borgonovi (2008: 2331).

[23] If we allow you to drop out from *Option 2* after several years, doesn't this introduce a possibility that making the relevant commitment doesn't promote extreme good, as premise (1) claimed? I don't think it does. Given the numbers referenced in fn.7, it seems that even a single year spent in advocacy for effective charities should qualify as promoting extreme levels of good in expectation.

deal? Perhaps by causing you to miss vital training opportunities and thus costing you a maximally competitive career in (say) some other sector in which you had been interested, such as the legal or banking sector?

A great deal will obviously depend upon the details. For a start, though, it's not obvious that having a maximally competitive or maximally pecuniary career is the sort of thing which generally makes your life go *better* – especially given some of the surprisingly weak correlations between income and life satisfaction I mentioned previously. Overall, I do wish to resist the idea that sacrificing (say) two years of your life towards paths like *Option 2* would negatively impact your lifetime welfare to a substantial degree. Apart from anything else, it is generally prudentially valuable to cultivate an open mind, to seek out a variety of experiences, to interact with a wide range of social groups, and so on, not least because these are the sorts of things which equip persons to know themselves and to know the sort of life they would like to carve out for themselves. These are distinctive benefits accrued by the agent in the event that they make a serious and concerted effort to *try* and make unorthodox lifeplans like *Option 2* work, even if they eventually decide that such an effort is unsustainable.

5. Premise (3)

Most of the heavy lifting is now done. I've argued that committing to an altruistically-focussed lifeplan is generally better for others and generally not expectedly worse for oneself. The final premise of the Lifeplan Argument, premise (3), makes the following claim: if committing to a lifeplan promotes extreme expected good for others without being expectedly worse for the one who commits to it, one is morally required to commit to that lifeplan. This premise can be seen as a specific instance of a much more general principle: all else equal, if an act *X* promotes extreme expected good for others without

being expectedly worse for the one who performs it, then one is morally required to perform *X*.

This general principle, I think, is especially tempting. In particular, there's at least one sense in which this general principle is weaker than the analogous principle used in Peter Singer's famous argument that each of us is morally required to engage in substantial levels of charitable giving.[24] Singer's general principle, readers may recall, was that if we can prevent something terribly bad from happening without sacrificing anything of comparable moral importance, then we are morally required to do so. Famously, an application of this general principle might go as follows: I am required to sacrifice my beloved expensive suit (thus making myself non-trivially worse-off) in order to save several children from drowning in a nearby pond. Note, however, that the analogous general principle underlying premise (3) is not even *that* strong. The analogous general principle underlying premise (3) remains silent about whether one would be required to sacrifice one's suit (and thus make oneself non-trivially worse off) in order to save several nearby drowning children. It instead carries the following sort of modest implication: if I can take two different paths to work, neither of which is worse for me than the other, then I am morally required to walk down the path which allows me to costlessly save several children from drowning in a nearby pond. If we found the general principle underlying Singer's argument tempting, we will want to uphold the general principle underlying premise (3) even more so.

Of course, whether or not we find premise (3) plausible is inevitably going to depend upon our underlying account of how duties of beneficence are generated in the first place. For instance, those who deny that we can ever have positive duties of beneficence towards

[24] Singer (1972: 231).

distant strangers will be tempted to simply reject (3) outright.[25] But it is easy to see why certain existing and intuitively appealing frameworks would lend support to premise (3). For instance, according to a popular framework with roots reaching back to Rossian pluralism, the fact that an act *X* would prevent severe harm (such as the painful and premature death of several persons) serves as a *requiring reason* in favour of *X* – that is, a reason which 'pulls in the direction' of *X* being required.[26] By contrast, the fact that an act *X* is costly for the one who performs it might generate a cost-based *permitting reason* against the performance of *X* – that is, a reason which pulls in the direction of *X* being merely permitted rather than required. Whether or not *X* is required is then going to depend upon a tug-of-war between the requiring versus permitting reasons at play in our evaluation of *X*.

It's not hard to see how the general framework of requiring versus permitting reasons might lend support to premise (3). There are requiring reasons in favour of committing towards altruistically-focussed lifeplans (given that such commitments promote extreme expected good for others). There are not cost-based permitting reasons against committing towards altruistically-focussed lifeplans (given that such commitments are not expectedly worse for oneself). What thus begins to emerge, given the balance of requiring versus permitting moral reasons, is a moral requirement to pursue altruistically-focussed lifeplans.[27]

If we accept premises (1)–(3), then we arrive at the central conclusion of interest. We are each required to begin our moral lives by committing ourselves towards an

[25] Perhaps this is broadly Narveson's (2004: 403) view.
[26] Ross (1930). The imagery of "pulling towards" is from Pummer (2023: 24–25), who provides the most comprehensive recent articulation of the requiring versus permitting reasons framework. N.B. Pummer (2023; 2025) allows for more than cost-based permitting reasons, including autonomy-based permitting reasons, which might cut against requirements of beneficence. I revisit the important complication posed by considerations of moral autonomy in §7–8.
[27] Again, assuming that these reasons were to be exhaustive, which Pummer (2023; 2025) denies; a complication to be revisited shortly.

altruistically-focussed lifeplan: that is, a lifeplan devoted in a radical and consuming way towards the promotion of benefits for other persons. Those of us who have failed to do so have failed to do our moral duty.

6. Practical Implications

Momentarily, I'll turn to consider potential objections. First, though, it may be worth setting out some important practical implications that the Lifeplan Argument brings with it, including potential implications for persons *not* at the start of their moral lives.

The first reason that the Lifeplan Argument offers to be of practical importance, of course, is that it directly bears upon the moral responsibilities held by those of us at the start of our adult moral lives. Those persons who find themselves at the start of their adult moral lives are governed by a moral requirement to commit to making the rest of their lives unfold in a manner which is 'altruistically-focussed.' This, it goes without saying, is a highly revisionist account of how persons must act when deliberating about the structure of their future.

A natural worry is that this central result only bears upon the moral responsibilities of a select few persons – namely, persons at the very beginning of their adult moral lives. What I want to emphasise here, however, that this central result also seems to yield important implications for persons who are *not* at the start of their adult moral lives.

Suppose, for instance, that I am midway through my adult moral life – I am thoroughly enraptured in certain projects, endeavours, and commitments which I collectively view to imbue my life with a sense of meaning and purpose. Since I am already deeply enraptured in these projects, giving them up would indeed be extremely costly – giving up my artistic *magnum opus*, perhaps, would emotionally devastate me for a period of one year.

What seems plausible, here, is that whether or not I have a moral permission to continue pursuing these personal projects depends (at least in part) upon whether I had a moral permission to seek out a lifeplan containing those personal projects in the first place. Suppose, for instance, that I *did not* have a moral permission to embark upon a lifeplan containing those artistic projects in the first place. If this is the case, it's tempting to think that I may be required to give up those artistic projects for the sake of others, even if doing so turns out to be highly personally costly – even if, in normal circumstances, we would reject such a requirement as overly burdensome.[28] The general principle at work, here, seems to be something like the following: it is easier to generate a moral requirement that we abandon even very wonderful personal goods and projects in the event that we did not have moral entitlements to seek out those personal goods and projects in the first place.

It would be dubious, I think, to argue that one is *always* required to give up those projects upon which one wrongfully embarked in the first place. To take an independent case: imagine that you wrongfully initiated an experimental medical research project which unjustifiably harmed patients at some previous point in time t_1. You now find yourself at t_2. The unjustified harms have already been done. Furthermore, completing the medical project will unlock important medical advances. In this case, perhaps it is permissible (or even required) that you complete your medical research project. A similar point could be true of certain prudential projects, such as the aforementioned artistic *magnum opus*. Even if I did not have a moral permission to embark upon my artistic *magnum opus*, perhaps, being extremely close to the point of completion, I might still hold a moral permission to finish it now.

[28] Sentiments explored at greatest depth in Barry & Øverland (2016) and Cordelli (2018). In particular, Barry & Øverland (2016: 30–49) draw attention to the fact that the size of your historical 'assistance shortfall' seems relevant in measuring the strength of your current *additional* obligations, where an assistance shortfall constitutes the degree to which you previously fell short of sacrificing minimally required costs for the sake of others.

Nonetheless, even in such cases, the fact that the project was *originally* prohibited can still be of moral relevance. I take one common reaction to the medical research case to be something like the following: although the medical project may be finished, it may only be finished under certain stringent conditions. Perhaps the research project must be accomplished by a public awareness campaign, or by a system of social cues such that it is impossible to receive benefits produced by the medical program without simultaneously being made aware of the historical injustices which brought about those benefits. Similar things may be true of the artist who is on the verge of completing her *magnum opus*. Perhaps the *magnum opus* may still be completed. But perhaps the *magnum opus* must be completed under certain stringent conditions. Perhaps, for instance, one is required to give away a substantial portion of the artistic project's financial proceeds, or accompany the artistic project with certain kinds of social cues which encourage charitable giving.

In these two ways, then, the Lifeplan Argument offers to illuminate the obligations of beneficence held by moral agents not at the start of their moral lives. It becomes easier to generate requirements that agents *give up* projects upon which they wrongfully commenced. And it becomes easier to generate requirements that agents *change the manner* in which they conduct projects upon which they wrongfully commenced.

It is hard to overstate the importance of these conclusions, I think, for the existing philosophical literature on beneficence. For the last forty or so years, philosophers have spilled a great deal of ink arguing over what is an intuitively reasonable proportion of one's disposable income to donate towards charitable causes.[29] Various principles such as act utilitarianism have been rejected on the grounds that they require us to donate very large portions of our disposable income towards charitable causes and on the grounds that such a requirement is inappropriately demanding. The Lifeplan Argument, however,

[29] As in fn.2.

lends very high credence to the possibility that we *are indeed* regularly required to donate very large portions of our disposable income towards charitable causes. And it lends credence to this possibility regardless of whether or not we are attracted to (say) act utilitarianism in particular. It lends credence to this possibility by drawing attention to the fact that we often lacked a prior moral right to seek out this-or-that project, associated with this-or-that disposable income, in the first place.

7. Replies

My final task is to respond to potential objections. In particular, many non-consequentialists may remain suspicious of my premise (3): the premise that one must commit towards lifeplans which are better for others and no worse for oneself. Such non-consequentialists may worry that this premise neglects the importance of important non-consequentialist concepts – concepts such as partiality, integrity, psychological difficulty, and autonomy. In what follows, I offer a conciliatory response. I argue that the Lifeplan Argument is indeed compatible with a robust concern for each such non-consequentialist consideration.

7.1 Partiality & Integrity

Premise (3) of the Lifeplan Argument claims that we are required to commit towards lifeplans which are much better for others and no worse for ourselves. But one might object that this requirement neglects the importance of the agent's *current partial projects* in a manner which is somehow ethically inappropriate.[30]

[30] The *locus classicus* being Scheffler (1982). The current concern could be glossed as an objection to premise (2) if one felt that being required to 'forfeit the current partial perspective' somehow caused persons' lives to go especially poorly.

It may already be clear that this reply faces special difficulties in the current context. Since the Lifeplan Argument considers agents at the start of their moral lives, the agents in question lack – at least in *some* distinctive way – those partial projects which the relevant commitment would have otherwise undermined. The agents in question lack these partial projects precisely because they are in the process of choosing which partial projects *are the ones to which they will commit themselves*. In light of this fact, it becomes much harder to argue that the relevant requirement problematically neglects the importance of one's partial projects.[31]

This is too hasty, perhaps. I've already emphasised that each of us come into the world already wrapped up in partial relationships, as with our parents or siblings. And many of us have things which can be appropriately characterised as partial projects even from a very young age. Wouldn't the Lifeplan Argument still neglect the importance of these early attachments?

I don't think so. As for partial relationships: nothing I have said so far about the nature of the altruistically-focussed life would require cutting-off (say) pre-existing partial relationships formed in infancy. What would instead be required, presumably, is situating or integrating these early relationships within a life with an overall altruistic focus (as paradigm moral saints like Paul Farmer often do). If this is right, it is not so obvious that the relevant requirement can be accused of undermining one's early partial relationships.

Might the relevant requirement nevertheless undermine one's early partial projects? Again, I'm not sure it would. One important possibility is that, in general, our attachment to our partial projects regularly conforms to a *chain link* structure. The chain link structure is represented in the following diagram, where the x-axis represents time and

[31] This riposte is not new; similar points are made in Slote (1985: 23–34).

the *y*-axis represents the intensity of one's attachment to any given project *p* at a given point in time:

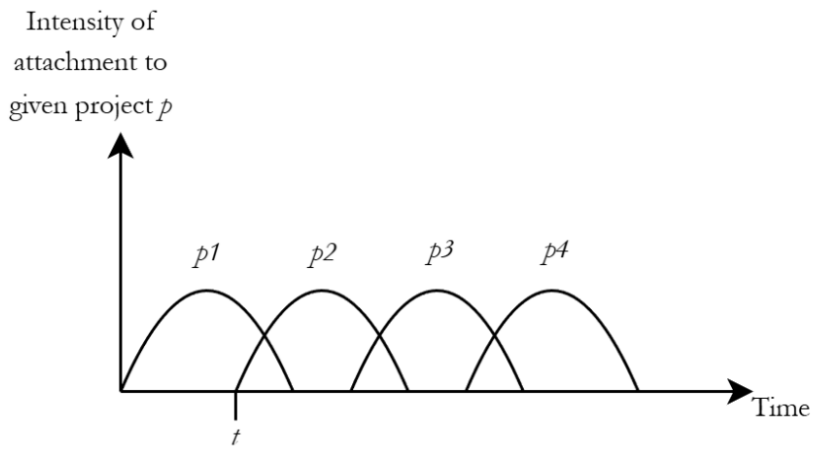


Fig.1 Chain Link Structure of Project Attachment

The overlap between different partial projects is representative of the fact that life is messy, and that each of us is often torn between several partial projects at any given time at different stages of completion. The important point to make here is that agents at the sort of 'lifeplan choice points' I've so far considered can often be understood as occupying the point in time *t*. Even though the agent might be committed to certain early partial projects, she nonetheless faces an impending choice about what sort of *new* projects she will initiate. The Lifeplan Argument then claims that there is a restriction upon the sorts of new projects upon which the agent may legitimately embark. Insofar as it does so, it's much harder to argue that the Lifeplan Argument problematically neglects the importance of one's partial projects.

Let me briefly add that, for a very similar set of reasons, I am unconvinced whether Bernard Williams' famous notion of *integrity* would be of help in resisting the argument.[32]

[32] Williams (1973).

Of course, it's notoriously difficult to interpret the notion of integrity.[33] Centrally, the idea seems to be that an agent should be allowed to navigate her life while preserving and respecting those goods and projects which *she herself* understands to be of genuine importance: that the agent cannot be required to forfeit *her* goods and projects in a manner which undermined her sense that she was still a unified person, someone who held her own set of values and her own distinctive perspective upon the world. I think such an objection, too, would lack dialectical force in the current context. Since we are considering agents at the start of their moral lives, the relevant agents are – at least in *some* distinctive way – choosing those goods and projects the very preservation of which will turn out to constitute their integrity.[34]

7.2 Psychological Difficulty

Premise (3) of the Lifeplan Argument claims that we are required to commit towards lifeplans which are much better for others and no worse for ourselves. But one might object that this neglects the importance of *psychological difficulty* in a manner which is somehow ethically inappropriate.[35]

For instance, according to the pluralist framework recently developed by Brian McElwee, a requirement can sometimes be rejected as overly demanding insofar as (i) that requirement is overall costly to welfare, or insofar as (ii) it is *psychologically difficult* to conform with the requirement in question.[36] Roughly speaking, the notion of 'psychological difficulty' attempts to convey that the act in question requires a strenuous

[33] See Chappell (2007) for some of the possibilities.
[34] By contrast, if the notion of 'integrity' is glossed as something akin to a right to self-authorship over time, or a right to authentic self-exploration, I deal with related challenges in §8.
[35] Chappell (2019), McElwee (2023). C.f. van Ackeren (2018).
[36] McElwee (2023).

exertion of willpower to perform.[37] The proposed view, here, is that the presence of psychological difficulty is not *per se* a reduction in welfare itself, but rather, a non-welfarist mental phenomenon capable of moderating obligations.[38]

Proponents of the psychological difficulty view could grant that altruistically-focussed lifeplans are better for others and no worse for me. But they might then reject any proposed requirement that we commit ourselves to those lifeplans. They might reject such a proposed requirement on the grounds that it is overly demanding in the psychological difficulty sense.

Again, it may already be clear that this reply faces some special difficulties in the current context. There are special reasons for believing that the requirement specified by the Lifeplan Argument may not be especially psychologically difficult – at least, not psychologically difficult to the extent that the requirement could be rejected as all-things-considered overly demanding.

An initial observation is that, as we've just seen, the Lifeplan Argument generally requires that we give up certain *opportunities* to form particular partial projects rather than giving up partial projects which we have already formed. It can, of course, be psychologically hard to give up opportunities. But it is often much psychologically easier to give up an opportunity 'out in the ether' than it is to give up an particular project or attachment upon which one has already embarked. One general explanation for this is that mere opportunities can often be given up without the exercise of active volition: I can give up on an opportunity to go for a morning walk by simply continuing to sit in bed.

It's similarly important to bear in mind, here, some of my previous remarks about *trying*. What will often make the relevant commitment much psychologically easier is that

[37] McElwee (2023: 14–15).
[38] Of course, psychological difficulty can be qualitatively unpleasant in ways which might be relevant to welfare; for discussion of cases and complications, see McElwee (2023: 6–8).

the commitment in question does not require the agent to irrevocably commit herself to a given altruistic lifestyle for perpetuity. It only requires that the agent make a serious and intentional commitment to *try* and bring about an altruistically-focussed lifestyle, where such a requirement is compatible with a *post hoc* decision to leave that lifestyle. Again, this is the sort of fact which should make the relevant requirement psychologically easier.

One final point worth emphasising goes as follows. Note that, insofar as we step back and take an aggregative perspective, it is deeply unclear whether there is greater psychological difficulty involved in actively living out a life like *Option 2* than there is in actively living out a life like *Option 1*. It is unclear, for instance, whether moral saints like Paul Farmer are easily characterised as being wracked by the constant experience of willpower strain. By contrast, it is plausible that various 'conventionally prudential' lifeplans centred around (say) the meeting of professional goalposts contain tremendous amounts of psychological difficulty on a regular basis. Insofar as we think that the aggregative perspective is relevant for the evaluation of psychological difficulty, then, it isn't obvious that altruistically-focussed lives like *Option 2* qualify as containing greater psychological difficulty overall.

7.3 Moral Autonomy$_2$

Premise (3) of the Lifeplan Argument claims that we are required to commit towards lifeplans which are much better for others and no worse for ourselves. But one might object that this neglects the importance of *moral autonomy* in a manner which is somehow ethically inappropriate.

Unfortunately, there are many different things we can mean in using the word 'autonomy,' let alone in using the contested phrase 'moral autonomy.'[39] Influentially, Seana Shiffrin drew attention to two different kinds of moral autonomy which might stand in need of safeguarding.[40] The first kind of moral autonomy, *moral autonomy*$_1$, is secured when an agent has the ability to engage in uncoerced deliberation about what is morally required of her and to translate these deliberations into actions. A moral agent with *moral autonomy*$_2$, by contrast, routinely possesses 'moral breathing room': they routinely possess the ability to select between multiple permissible courses of action. As Shiffrin puts it, moral autonomy$_2$ centrally involves an agent's:

> … ability to make an autonomous choice from a range of morally permissible options. If one has moral autonomy$_2$, then one has the opportunity to make a decision between at least two alternatives where whichever choice one makes, one will not be acting impermissibly.[41]

These two varieties of moral autonomy, suggests Shiffrin, can come apart. And we can sometimes reject moral rules or standards, suggests Shiffrin, on the grounds that those rules or standards provide persons with an insufficient degree of moral autonomy$_2$.[42]

You might think that, with the tool of moral autonomy$_2$ in our belt, we are in a position to reject premise (3) of the Lifeplan Argument. You are *not* required to start your life by committing towards an altruistically-focussed lifeplan, even if doing so is better for others and not worse for yourself. For such a requirement is incompatible with you routinely having and exercising moral autonomy$_2$ – 'breathing room' – of the kind indicated here.

---

[39] See Arpaly's (2004) taxonomy.
[40] Shiffrin (1991). Shiffrin follows earlier remarks in Slote (1985), and Shiffrin positions her remarks as a potential interpretation of Slote which would be inoculated from criticisms in Kagan (1989). It is a common criticism of Slote (1985), as in Howard-Snyder (1992), that Slote underspecifies his exact conception of moral autonomy. In light of this fact I focus upon Shiffrin's interpretation. For recent discussion, see Lazar (2019), Pummer (2023; 2025), and Tucker (n.d.).
[41] Shiffrin (1991: 246).
[42] Shiffrin (1991). Pummer (2023: 28) follows Shiffrin in emphasising the importance of breathing space.

Upon reflection, however, it seems that moral autonomy$_2$ cannot ground a successful objection to premise (3) of the Lifeplan Argument. The most obvious reason for this is that one can be determinately required to select a given *type* of campaign of action without it being the case that, at each point within that campaign of action, one has only a single and uniquely permissible option. For instance, in **Frost's Choice**, there might well be a moral requirement that I pursue the life of malarial doctoring. But a requirement to embark upon such a campaign may simply remain silent with regards to the question of what I should have for breakfast next week. It may also be silent on more substantive questions such as the order of operations which my career preparations should take, the exact school at which I should study, what I should do with my spare moments, … and so on. If this is right, then the relevant requirement specified by the Lifeplan Argument is entirely compatible with the regular presence of robust moral autonomy$_2$ – that is, with the regular presence of moral breathing space.

Perhaps the response from moral autonomy$_2$ would have had traction if I had been advocating a utilitarian requirement that agents must pursue *exactly* those series of steps conducive to the maximisation of expected utility. Such a requirement might indeed be incompatible with the species of moral autonomy$_2$ pinpointed by Shiffrin. But I have not been advocating for any such requirement: the Lifeplan Argument only argues that, at the start of their moral lives, agents must make a serious and intentional commitment towards a lifeplan with an altruistic focus.

8. Compositional Autonomy

Of the above replies, the reply from moral autonomy strikes me as the most promising. I'll thus conclude by pushing things a little further; by considering whether the reply from moral autonomy can be workshopped into something more plausible. In particular, I'll

consider the following suggestion: premise (3) of the Lifeplan Argument is incompatible with a robust concern for (what we might call) the *compositional autonomy* of individual decision makers.

8.1 Compositional Autonomy

Suppose that premise (3) of the Lifeplan Argument were correct. It would then follow that you were determinately required to select *Option 2* in **Frost's Choice.** You would be determinately required to pursue the career of malarial doctoring.

Now, as I've suggested, this requirement seems compatible with a robust level of moral autonomy$_2$ – you might regularly have 'moral breathing room' to choose between (say) different breakfast cereals or different evening pastimes. Nevertheless, this requirement seems like it might be incompatible with a different kind of moral autonomy. It might be incompatible with you having (as I'll call it) *compositional moral autonomy:* the moral freedom to freely and creatively make the core of your future life to be about some contrastively different domains of value rather than others. For instance: you would be furnished with a moral permission to make the core of your future life to be *about malarial vaccination.* But you wouldn't be furnished with a moral permission to make the core of your future life to be about *anything* else.[43]

What do I mean, in the above definition, in saying that compositional autonomy consists of being able to make the core of your future life to be about 'contrastively different domains of value'? For now, let's just note that there are two obvious ways in which a decision maker might encounter a choice between contrastively different domains

[43] Of course, in emphasising the importance of moral autonomy$_2$, Shiffrin might be attempting to emphasise that it is important for agents to possess compositional autonomy of the kind alluded to here. Certain remarks at the close of Shiffrin's (1991: 251–252) discussion gesture in this direction. If so, then the following might be seen as a potential interpretation or elucidation of Shiffrin's remarks.

of value: they might encounter a choice between moral and non-moral values ("Prevent a lethal car crash or restore *The Night Watch"*) or alternatively between contrastively different moral values ("Prevent a lethal car crash or improve child nutrition within your community"). Both kinds of contrast will matter, presumably, insofar as we are asking about the extent to which someone possesses compositional autonomy with respect to their future life.[44]

What do I mean, in the above definition, in saying that compositional autonomy consists of being free to put contrastively different values at the 'core of one's future life'? If we can intuitively grasp how idea works in canonical cases like **Frost's Choice**, then I'm not sure we need a fully precise analysis. But one answer, which strikes me as having as much plausibility as any other, is that I make a value into the core of my future life when that value becomes a component of my *individual identity*: that is, when that value becomes one of the primary facts which helps in giving an account or explanation to others of who I am. Political and religious values are, in this sense, more core to one's individual identity than are incidental opinions about the weather. Such values play a role in my attempted answer when other persons ask of me: "Who is it that you *actually are*?"[45]

One might reject the premise (3) of the Lifeplan Argument on precisely on the grounds that it is incompatible with humans having compositional autonomy of the kind sketched above. Human beings *clearly do* have compositional autonomy with regards to their future lives: human beings clearly *do* have a kind of moral freedom to make the core of their

[44] For current purposes we can simply assume that for a good x to be valuable is for x to be sort of object about which positively valenced pro-attitudes are appropriately held. For a survey of fitting-attitude analyses, see Rabinowicz (2013).

[45] Naturally the concept of individual identity is contested. For an appealing account see Schechtman (1996). For a sceptical treatment see Mogensen (2021).

future life to be about various contrastively different domains of value. Since premise (3) tries to deny as much, the thought goes, premise (3) ought to be rejected.[46]

8.2 Compositional Autonomy and Its Siblings

Let me very briefly emphasise that this way of replying to the Lifeplan Argument wouldn't be entirely novel. It's been a recently popular sentiment that considerations of moral autonomy must *somehow* be capable of blocking a requirement to doctor in cases like **Frost's Choice**, even if the details sometimes prove difficult.[47]

For instance, I initially suggested that premise (3) of the Lifeplan Argument gains plausibility under the requiring versus permitting reasons framework: I observed that there are requiring reasons in favour of performing *Option 2* but no cost-based permitting reasons against performing *Option 2*. But Pummer, in his recent defence of the requiring versus permitting reasons framework, is keen to emphasise a further category of *autonomy-based permitting reasons*; a further type of permitting reason which generates moral permissions in favour of performing sub-optimific actions.[48] For Pummer, the fact that an act of mine φ would secure a *distinctive valuable feature x* in my future life generates an autonomy-based permitting reason in φ's favour, that is, a permitting reason which pulls in the direction of φ being morally permissible. With respect to **Frost's Choice**: the fact that *Option 1* secures so many distinctive valuable features would generate strong

[46] N.B. There are potential relationships between compositional autonomy and other ethically influential concepts. The most notable related concept may be the ethical ideal of *authenticity*, as articulated in Taylor (1991) and Williams (1993: 79; 2002: 172–205). The ethical ideal of authenticity is generally thought to be achieved when persons follow D. H. Lawrence's (1923) maxim, "Find your deepest issue, in every confusion, and abide by that." The ideal in question generally presupposes that there exists a 'true self' and that ethical life should consist, at its core, of attempts to uncover and conform to that true self (perhaps within suitably specified constraints, as in Taylor (1991: 31–41)). It's unclear to me that the requirement to try altruistic lifeplans like *Option 2* is anathema to the pursuit of authenticity, since one way of uncovering aspects of one's 'true self', if there is such a thing, is to try and embark upon unorthodox lifestyles such as *Option 2*.
[47] See fn.40.
[48] Pummer (2023; 2025).

autonomy-based permitting reasons in its favour, where these reasons are likely to render *Option 1* permissible.[49]

Similarly, according to the account of moral reasons recently defended by Tucker, valuable non-moral goods secured by some act φ can have a high level of *justifying weight* with respect to φ – that is, weight which pulls towards φ being rendered permissible – because non-moral goods (such as aesthetic ones) have special fit for the project of autonomous "self-making."[50]

My above observation that premise (3) of the Lifeplan Argument would be incompatible with compositional autonomy is of a kind with these recent suggestions. Of course, there are differences in the details. One reason why it may be useful to employ the notion of compositional autonomy, at least for present purposes, is that this concept is a slightly more ecumenical one: Pummer's articulation of autonomy-based permitting reasons, for instance, is developed firmly within the particular framework of requiring versus permitting reasons. One could, in principle, emphasise the importance of compositional autonomy in cases like **Frost's Choice** without committing to the details of such a reasons framework.

8.3 *A Conciliatory Response*

At last, then, it seems like we might have a promising objection to premise (3) of the Lifeplan Argument: premise (3) of the Lifeplan Argument must be false because it is incompatible with the existence of compositional autonomy. I think, however, that even the objection from compositional autonomy can be overcome. Indeed, as things seem to me, a slightly modified version of the Lifeplan Argument would nonetheless prove

[49] For recent criticism of Pummer (2025), see Tucker (n.d.).
[50] Tucker (n.d.: 14–16).

expressly compatible with a robust concern for considerations of compositional autonomy.[51]

The key suggestion goes as follow. Perhaps, if you really *did* find yourself in the binary and highly idealised **Frost's Choice**, you would be permitted to select the career of architecture rather than the career of malarial doctoring. Perhaps considerations of 'compositional autonomy' would demand as much.

Nonetheless, this is not the situation which many of us face in the actual world. In the actual world, many of us begin our moral lives faced with a *large range* of potentially viable career plans or lifeplan paths, where many of these potentially viable career plans or lifeplan paths can be appropriately characterised as 'altruistically-focussed' in the sense I have sketched. (Think, here, of careers in malarial doctoring, of careers in AI safety, of careers in charity fundraising, … and so on.)

We might think that, in general, if one is faced with a large range of altruistically-focussed lifeplans which are better for others and no worse for oneself, then one is required to commit oneself to one of those lifeplans – even if there is no altruistically-focussed lifeplan *in particular* to which one is required to pre-commit oneself. Distilling these sentiments, we would arrive at the:

[51] Of course, a more hardline response would involve simply following Kagan (1989: 237–238) in denying that moral autonomy is a coherent or important concept. Here, I am interested in exploring a more conciliatory reply.

Qualified Lifeplan Argument
(1*) At the start of their adult moral life, agents face a large set of altruistically-focussed lifeplans which promote extreme expected good for others.
(2*) At the start of their adult moral life, each of those altruistically-focussed lifeplans is not expectedly worse for the agent who pursues it.
(3*) If an agent faces a large set of available lifeplans, where each such lifeplan promotes extreme expected good for others without being expectedly worse for the agent who pursues it, she is morally required to commit to *one* of those lifeplans, even though she is not required to commit to any such lifeplan in particular.
(4*) Therefore, at the start of their adult moral life, agents are required to commit themselves to *one* altruistically-focussed lifeplan, even though they are not required to commit to any such altruistically-focussed lifeplan in particular.

The Qualified Lifeplan Argument would still demonstrate a highly revisionist conclusion. It would still demonstrate that each of us must begin our lives by committing ourselves to *an* altruistically-focussed lifeplan. It would simply allow that there is no *particular* altruistically-focussed lifeplan towards which any given person must pre-commit herself. Such a result would still depart substantially, in this way, from our commonsense moral appraisals.

Crucially, however, it would be extremely difficult to try and reject the Qualified Lifeplan Argument by citing considerations of compositional autonomy. Why? Well, the most obvious reason is that, under the Qualified Lifeplan Argument, agents are still provided with a substantial degree of moral liberty when deciding what sort of values will sit at the core of their future life. A requirement that the agent merely pursues *an* altruistically-focussed lifeplan leaves it open whether the agent should make the core of their future life to be about (for instance) the values involved in malarial doctoring, versus the values involved in AI safety research, versus the values involved in charity fundraising, … and so on. The Qualified Lifeplan Argument thus, in the end, seems to leave decision-makers with a very sizeable capacity for compositional autonomy.

Of course, the sceptic might be entirely unpersuaded by this retreat to the Qualified Lifeplan Argument. The sceptic might reply: "It would still be a *very substantial* reduction

to my compositional autonomy if I were required to restrict my attention to the set of altruistically-focussed lifeplans. For this reason, the Qualified Lifeplan Argument is still defective."

I don't find this sceptical reply especially convincing, however. Let me limit myself to three brief points, each of which is indicative of an underlying concern, perhaps, that a great deal would need to be said in order to fill in the details of this sceptical reply and thus make it persuasive.

First, it is very easy to undersell – as the sceptic may be doing – the degree to which different altruistically-focussed lifestyles can contain an enormous diversity of both moral *and* non-moral values. Think of some of the diversity which distinguishes, say, a career of malarial doctoring from a career researching AI safety from a career working as the marketing director of an effective charity. These careers aren't just diverse in the sorts of moral values they realise. They are also diverse in the sorts of *non-moral* values they realise. To be an effective malarial doctor requires the realisation of medical values. To be an effective technical researcher on AI safety requires the realisation of mathematical values. To be an effective marketing director within an effective charitable fundraiser requires the realisation of aesthetic values. And so on. One factor which explains *why* these altruistically-focussed lifestyles contain so many non-moral values is that the activities and behaviours which constitute these lifestyles are regularly 'multiple realisers' of both moral and non-moral values. To distribute malarial vaccines is, *ipso facto*, to promote both medical values and moral values.

Second, presumably the sceptic cannot be arguing for a kind of *absolutism* about compositional autonomy: arguing that any kind of (non-harmful) lifeplan must always remain permissible, and that considerations of compositional autonomy force us to accept as much. This view seems far too strong. Suppose, for instance, that members of some

country *A* can spare members of some country *B* from horrific death by collectively deciding to refrain from pursuing one very particular kind of lifeplan – lifeplans involving underwater basket weaving, say. Presumably, the *A*-members might be morally required to refraining from pursuing lifeplans containing underwater basket weaving, even though this restriction marginally diminishes their capacity for compositional autonomy.

What seems much more plausible than an absolutist model, I think, would be a kind of sufficientarian model: a model according to which the correct moral norms must leave persons with *enough* compositional autonomy when it comes to deciding what sorts of valuable goods their future lives will contain. But if this is right, the Qualified Lifeplan Argument still retains a very high level of plausibility. For the reasons I've sketched, the Qualified Lifeplan Argument still leaves persons with a very substantial capacity for compositional autonomy, even if their remaining capacity for compositional autonomy is not infinite.

The third and final point is familiar. In direct and immediate comparisons, we often feel that moral values should trump non-moral values in settling how we ought to act. As Singer and Unger observe, although there are immense values of engineering and craftsmanship distilled within Ferraris, you must crush your Ferrari if doing so would save a nearby child from an oncoming train.[52] The sceptic might indeed intuit, in response to the Qualified Lifeplan Argument, that we have an unrestricted moral permission to exclusively shape our future lives around non-moral values. But one possibility is that this intuition comes about because the suffering of others, in lifeplan choices like **Frost's Choice**, is far away on the distant horizons of space and time. Perhaps this intuition would vanish if we could bring ourselves to see that suffering more clearly.[53]

[52] Unger (1996), Singer (1999).
[53] C.f. Kamm (2000) and Slote (2007). For an extended recent treatment, see Pummer (2023).

9. Conclusion

Philosophers have written a great deal about beneficence as an obligation to donate one's disposable income to charity. We should be careful, however, not to miss the forest for the trees. In these pages I've argued that we are each governed by a much grander obligation of beneficence: an obligation to begin our adult moral lives by making serious and intentional commitments towards the living out of altruistically-focussed lifeplans. I've suggested that lifeplan choice points are distinguished by some interesting and unusual features which help us derive this conclusion. And I've suggested that this result sheds light on the obligations of beneficence we hold throughout our moral lives as a whole.

Although complete theory neutrality is impossible, I've argued for this conclusion by drawing upon a small number of ecumenical premises. Since the conclusion I've derived is so revisionist, the argument leaves us in an uncomfortable place. We must either accept a highly revisionist picture of our obligations of beneficence, or alternatively, we must jettison one of three seemingly appealing claims about the nature of charity. If these three claims are the ones which ought to be jettisoned, then perhaps the above discussion can progress our understanding of charity in other ways too.

References

van Ackeren, M. (2018). 'How Morality Becomes Demanding Cost vs. Difficulty and Restriction', *International Journal of Philosophical Studies*, 26(3): 315–334.

Aknin, L., Barrington-Leigh, C. P., Dunn, E. W., Helliwell, J. F., Biswas-Diener, R., Kemeza, I., Nyende, P., Ashton-James, C. E., & Norton, M. I. (2013) 'Prosocial Spending and Well-being', *Journal of Personality and Social Psychology*, 104(4): 635–652.

Arpaly, N. (2004). 'Which Autonomy?', in J. Campbell, M. O'Rourke, & D. Shier (eds), *Freedom and Determinism*. MIT Press.

Ashford, E. (2003) 'The Demandingness of Scanlon's Contractualism', *Ethics*, 113(2): 273–302.

Barry, C. & Øverland, G. (2016) *Responding to Global Poverty: Harm, Agency, and Responsibility*. Cambridge University Press.

Borgonovi, F. (2008) 'Doing Well by Doing Good. The Relationship Between Formal Volunteering and Self-Reported Health and Happiness', *Social Science and Medicine*, 66(11): 2321–2334.

Broome, J. (2022) 'Incommensurateness is Vagueness' in H. Andersson & A. Herlitz (eds), *Value Incommensurability: Ethics, Risk, and Decision Making*. Routledge.

Carbonell, V. (2009) 'What Moral Saints Look Like', *Canadian Journal of Philosophy*, 39(3): 371–398.

——— (2012) 'The Ratcheting-Up Effect', *Pacific Philosophical Quarterly*, 93(2): 228–254.

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

Chappell, R. Y. (2019) 'Willpower Satisficing', *Noûs*, 53(2): 251–265.

Chappell, T. (2007) 'Integrity and Demandingness', *Ethical Theory and Moral Practice*, 10(3): 255–265.

Cordelli, C. (2018) 'Prospective Duties and the Demands of Beneficence', *Ethics*, 128(2): 373–401.

Crisp, R. (2006) 'Hedonism Reconsidered', *Philosophy and Phenomenological Research*, 73(3): 619–645.

Cullity, G. (2003) 'Asking Too Much', *The Monist*, 86(3): 402–418.

——— (2004) *The Moral Demands of Affluence*. Oxford University Press.

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

Hooker, B. (2000) *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*. Oxford University Press.

Howard-Snyder, F. (1992) 'Elbow-Room for Consequentialists', *Analysis*, 52(4): 249–253.

Hui, B, Ng, J, Berzaghi, E, Cunningham-Amos, L & Kogan, A. (2020) 'Rewards of Kindness? A Meta-Analysis of the Link Between Prosociality and Well-Being', *Psychological Bulletin*, 146(12): 1084–1116.

Kagan, S. (1989) *The Limits of Morality*. Oxford University Press.

——— (2011) 'Do I Make a Difference?' *Philosophy & Public Affairs*, 39(2): 105–141.

Kamm, F. (2000) 'Does Distance Matter Morally to the Duty of Rescue?', *Law and Philosophy*, 19(6): 655–681.

Kidder, T. (2003) *Mountains Beyond Mountains*. Random House.

Lawrence, D. H. (1923) *Studies in Classic American Literature*. Thomas Seltzer.

Lazar, S. (2019) 'Moral Status and Agent-Centred Options', *Utilitas*, 31(1): 83–105.

MacAskill, W., Mogensen, A., & Ord, T. (2018) 'Giving Isn't Demanding', in P. Woodruff (ed), *The Ethics of Giving: Philosophers' Perspectives on Philanthropy*. Oxford University Press.

McElwee, B. (2023) 'Cost and Psychological Difficulty: Two Aspects of Demandingness', *Australasian Journal of Philosophy*, 101(4): 920–935.
Mill, J. S. (1998). *Utilitarianism*. Edited by R. Crisp. Oxford University Press.
Mogensen, A. (2021) 'Is Identity Illusory?', *European Journal of Philosophy*, 29(1): 55–73.
Mulgan, T. (2001) *The Demands of Consequentialism*. Oxford University Press.
Murphy, L. (2000) *Moral Demands in Nonideal Theory*. Oxford University Press.
Narveson, J. (2004) 'Is World Poverty a Problem for the Wealthy?', *The Journal of Ethics*, 8(4): 397–408.
Noggle, R. (2009) 'Give Till it Hurts? Beneficence, Imperfect Duties, and a Moderate Response to the Aid Question', *Journal of Social Philosophy*, 40(1): 1–16.
Norcross, A. (2004). 'Puppies, Pigs, and People', *Philosophical Perspectives*, 18: 229–245.
Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.
Pummer, T. (2016) 'Whether and Where to Give', *Philosophy & Public Affairs*, 44(1): 77–95.
——— (2023) *The Rules of Rescue: Cost, Distance, and Effective Altruism*. Oxford University Press.
——— (2025) 'Compensated Altruism and Moral Autonomy', *Social Philosophy and Policy*, 42(1): 186–203.
Rabinowicz, W. (2013) 'Value, Fitting-Attitude Account of' in H. LaFollette (ed), *The International Encyclopedia of Ethics*. Wiley-Blackwell.
Ross, D. (1930) *The Right and the Good*. Oxford University Press.
Schechtman, M. (1996) *The Constitution of Selves*. Cornell University Press.
Scheffler, S. (1982) *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions*. Oxford University Press.
Shiffrin, S. (1991) 'Moral Autonomy and Agent-Centred Options', *Analysis*, 51(4): 244–254.
Singer, P. (1972) 'Famine, Affluence, and Morality', *Philosophy & Public Affairs*, 1(3): 229–243.
——— (1980) 'Utilitarianism and Vegetarianism', *Philosophy & Public Affairs*, 9(4): 325–337.
——— (1999) 'The Singer Solution to World Poverty', *The New York Times*, 5 September, 60–63.
Slote, M. (1985) *Common-Sense Morality and Consequentialism*. Routledge.
——— (2007) *The Ethics of Care and Empathy*. Routledge.
Stevenson, B., & Wolfers, J. (2013) 'Subjective Well-Being and Income: Is There Any Evidence of Satiation?', *American Economic Review*, 103(3): 598–604.
Taylor, C. (1989) *The Sources of the Self: The Making of the Modern Identity*. Harvard University Press.
Taylor, C. (1991) *The Ethics of Authenticity*. Harvard University Press.
Thomson, J. A. (2022) 'Relief from Rescue', *Philosophical Studies*, 179(4): 1221–1239.
Timmerman, T. (2015) 'Sometimes There Is Nothing Wrong With Letting A Child Drown', *Analysis*, 75(2): 204–212.
Tucker, C. (n.d.) 'Autonomy: How Lesser Goods Can Be Justifying Reasons', unpublished manuscript.
Unger, P. (1996) *Living High and Letting Die: Our Illusion of Innocence*. Oxford University Press.
Williams, B. (1973) 'A Critique of Utilitarianism', in J.J.C. Smart & B. Williams, *Utilitarianism: For and Against*. Cambridge University Press.
——— (1993) *Morality: An Introduction to Ethics*. Cambridge University Press.

——— (2002) *Truth and Truthfulness: An Essay in Genealogy*. Princeton University Press.
Wolf, S. (1982) 'Moral Saints', *The Journal of Philosophy*, 79(8): 419–439.
Woollard, F. (2015) *Doing and Allowing Harm*. Oxford University Press.

2.

Moral Autonomy and Lesser Numbers Permissions

Abstract. Many philosophers affirm the importance of *moral autonomy*. Such philosophers affirm, for instance, that we have a protected moral permission to pursue a lifelong career of flute-playing rather than a lifelong career of doctoring, even if doctoring is better for others and just as good for me. In this article I argue that, across a wide range of practically relevant rescue cases, moral autonomy theorists are committed towards belief in sizable lesser numbers permissions. There can be reasonable disagreement about how to interpret this result. But my own sense is that, since lesser number permissions of this sort are implausible, we should therefore consider rejecting the notion of moral autonomy.

1. Introduction

Let's begin with a familiar island-based dilemma:

> **Island.** Complete stranger Alice is dying on one island. Complete strangers Blaise, Charlie, Daniella, Ethan, and Francis are dying on a different island. You only have time to visit one such island. As a result, you can costlessly save some but not all.[1]

In **Island**, are you morally required to save the five in virtue of the fact that their group contains the greater number? Or, by contrast, do you have a moral permission to save merely Alice?

[1] Such cases have origins in Anscombe (1967), Foot (1967), and Taurek (1977).

Many philosophers believe that you are indeed morally required to save the five – more generally, that you are morally required to costlessly save the greater number. Say, of such philosophers, that they believe in greater numbers requirements (GNRs).[2]

A much smaller group of philosophers believe that it would nevertheless be morally permissible to save merely Alice – more generally, that you are morally permitted to costlessly save the lesser number. It has become common to refer to this smaller group of philosophers as 'numbers sceptics.'[3] I'm unsure if this moniker is especially helpful. I'll simply say, of such philosophers, that they believe in lesser numbers permissions (LNPs).[4]

In this article I show how widespread beliefs about the importance of *moral autonomy* would in fact entail LNPs within canonical rescue cases such as **Island**, where moral autonomy references, roughly, the commonsense suggestion that decision makers should be afforded a certain level of creative moral freedom when it comes to structuring the course and contents of their future lives. I'll argue that this entailment doesn't just hold of **Island**: it holds in a wide range of practically relevant rescue cases in which one can costlessly save the few or the many.

There are, of course, different ways of interpreting this result. One interpretation goes as follows: "Considerations of moral autonomy would systematically entail LNPs – so much the worse for considerations of moral autonomy, which we must therefore abandon!" The other interpretation goes as follows: "Considerations of moral autonomy would

[2] *Inter alia*, see Parfit (1978), Kavka (1979), Kamm (1985; 1993), Scanlon (1998: 232), Kumar (2001), and Hsieh, Strudler & Wasserman (2006). For recent defences see Pummer (2016), Sung (2022), Henning (2023), Barnett (2025).
[3] Following Otsuka (2004).
[4] Watershed discussions were Anscombe (1967) and Taurek (1977). Since then, see Munoz-Dardé (2005), Doggett (2013), and Setiya (2014). Several of the latter arguments attempt to vindicate lesser numbers permissions without making explicit use of Taurek's (1977) original premises, which remain controversial. For recent discussion of Taurek, see Muñoz (2022) and Barnett (2025). I omit, from subsequent discussion, the possibility that you have a permission to save none. The question is thus whether there exists a conditional obligation to save the greater number.

systematically entail LNPs – so much the better for LNPs, which we must therefore accept!"

I tentatively lean towards the former interpretation. That is, I find it very difficult to accept that there could be widespread LNPs in familiar cases such as **Island**. I am therefore tempted to take this result as a *reductio* of considerations of moral autonomy. But I won't try to dogmatically press this interpretation here. My main concern will be to carry out the important work of showing *that such an entailment exists*: of showing that one cannot be a moral autonomy theorist without simultaneously being committed towards belief in substantial LNPs in cases like **Island**.

The gameplan goes as follows. In §2 I provide a brief overview of what moral autonomy theorists typically believe. For the sake of clarity, I focus my attention on one particular framework: the framework of *autonomy-based permitting reasons* as developed in recent work by Theron Pummer.[5] In §3 I show how, across a wide range of practically relevant rescue cases, including **Island**, autonomy-based permitting reasons will in fact generate a permission to costlessly save the few rather than the many. This suggests that moral autonomy theorists are indeed committed to affirming LNPs in a wide range of practically relevant cases. In §4 I respond to potential objections, and in §5, I draw out some of important ways in which this 'autonomy argument' in favour of LNPs would differ from existing pro-LNP arguments.

§6 concludes. Again, to give away the punchline, I don't interpret the discussion as furnishing us with a knockdown proof in favour of LNPs. Rather, I interpret the discussion as raising a difficult dilemma. Either we must jettison some commonsensical belief about the importance of moral autonomy, or alternatively, reconcile ourselves to the existence of widespread LNPs.

[5] Pummer (2023; 2025).

2. Moral Autonomy

2.1 The Moral Autonomy Movement

As mentioned, my first task is to provide an overview of what moral autonomy theorists typically believe. The easiest way of doing so is to focus upon the following sort of case:

> **Magazine**. Flicking through a careers magazine, you discover that you could either become a professional flautist or a medical doctor. Each career will be equally prudentially good for you. However, the career of doctoring will be somewhat better for other persons: it will cure very many broken legs.[6]

In **Magazine**, what are you morally required to do?

Act utilitarianism generates the conclusion that, in **Magazine**, you are morally required to become a doctor. The reasoning here is fairly straightforward. Act utilitarians believe that we are morally required to perform those acts which bring about the greater total utility. But the career of doctoring, by assumption, promotes the greater total utility. After all, the career of doctoring is just as good for you and it is also much better for other persons. Act utilitarianism thus generates a requirement that you become a doctor.

Many philosophers find the straightforward act utilitarian analysis of **Magazine** rather unsatisfying. They can agree with the act utilitarian that it would be *good* if you pursued doctoring. But they baulk at the idea that you would be *morally required* to become a doctor – and in particular, at the idea that it would be morally wrongful to become a flautist. Such philosophers would much rather stick with the commonsense appraisal that both

[6] The careers of flute-playing and doctoring could be imprecisely equally good in virtue of containing parity goods, as in Chang (2002). For an extended defence of the idea that many altruistically-focussed lives are not in fact prudentially worse for those who seek them out, see my Chapter 1.

career paths are ethically permissible: that it is 'innocent' to go down any career path which does not actively do harm to others.[7]

In particular, a large number of *moral autonomy theorists* maintain that considerations of moral autonomy somehow generate a protected moral permission for you to pursue lifestyles such as "career flute-playing": more generally, a moral permission for you to pursue those lifestyles which are just as good for you but worse for others, or which are just as good for others but worse for you. What *is* moral autonomy, exactly? The concept remains difficult to pin down. But since influential discussions by Michael Slote, the general idea, in rough form, has been that persons should possess a certain kind of moral liberty to freely and creatively shape the course of their future life as they themselves would like to shape it, even if they would like to shape their future life in suboptimal ways.[8] A requirement to become a doctor in **Magazine** is incompatible with you and I possessing moral autonomy of this kind. Ergo, it cannot be the case that you and I are required to become doctors in **Magazine**.

Although numerous objections have been pressed against Slote's original discussion of moral autonomy, many philosophers continue to follow Slote in defending the importance of moral autonomy and in using the concept, or very closely related ones, to vindicate ordinary judgements in cases like **Magazine**.[9] One happy upshot of this fact has been the development of some concrete frameworks which might help explain *why* moral autonomy is an important or needed concept.

---

[7] Slote (1985: 26–27). N.B. Some utilitarian models, such as the varieties of satisficing utilitarianism discussed by Pettit & Slote (1984), may not carry the alleged implication that flute-playing is morally wrongful.

[8] Slote (1985: 27).

[9] For criticisms, see Kagan (1989: 237-238). Followers of Slote include Shiffrin (1991), Kamm (1992), Hurley (1995), Lazar (2019), Pummer (2023; 2025), and Tucker (n.d.). I explore one potential interpretation of the concept in Chapter 1's discussion of 'compositional autonomy.'

What are these frameworks? For current purposes, we need not pick between them. But the first framework, roughly speaking, would ground considerations of moral autonomy in considerations of *moral status.*[10] The reasoning goes as follows. As beings, you and I both hold unconditional value. Our holding unconditional value, it seems, is going to carry certain implications. One potential implication is that, in a case like **Magazine**, we cannot be required to pursue the option which is maximally conducive to the good of ourselves and others. For if we *were* required to pursue the maximally good-promoting option, it would not be appropriate to characterise us as rational and deliberative creatures who constituted ends in and of ourselves. It would be more appropriate, upon reflection, to characterise us as mere instruments or tools for the promotion of the good. Perhaps such an account of moral status must itself be taken as ethical bedrock. Nonetheless, drawing upon the idea of moral status might allow us to dig a little deeper down.

Another potential framework would ground considerations of moral autonomy in considerations of *belonging.*[11] The reasoning goes as follows. As beings, you and I own our respective bodies. Our bodies belong to us in the sense that you have a distinctive moral prerogative in settling how your body is to be used, and I have a distinctive moral prerogative in settling how my body is to be used. One potential implication of our bodies belonging to us is that, at least in certain cases like **Magazine**, we cannot be required to pursue those options which are maximally conducive to the good. For if we *were* always required to pursue those options which were maximally conducive to the good, there

[10] Explored at greatest length in Lazar (2019). Previous discussions, such as Kamm (1992), angle in the same direction by emphasising that it would be *valuable* (alternatively: good for welfare) for persons to have the capacity for options or moral autonomy. Such arguments might strike us as having the same unusual air that certain theological ontological arguments sometimes have: claiming that such a capacity must exist because it would be valuable (good for welfare) for such a capacity to exist. For a recent defence of this style of argumentation in ethics, see Preston-Roedder (2014).

[11] Pummer (2025) characterises this avenue as the *Rights View*: related ideas are explored in Quinn (1989) and Woollard (2013; 2015).

would be an insufficiently substantive sense in which our bodies genuinely belonged to us – in which we each possessed a distinctive moral prerogative in settling their respective use. Perhaps the notion of belonging must itself be taken as ethical bedrock. Again, however, drawing upon the idea of belonging might allow us to dig a little deeper down.

2.2 Autonomy-Based Permitting Reasons

The above is a fairly high-level sketch of the moral autonomy movement. For the purposes of subsequent discussion it will help to have a more precise target in mind. I'll take, as my target, the framework of *autonomy-based permitting reasons*, developed at length in recent contributions by Theron Pummer.[12]

Here is how the framework operates. Suppose that I am choosing between two acts φ and ψ, and that φ would spare someone from grave harm. According to the requiring versus permitting reasons framework, the fact that φ would spare someone from grave harm serves as a *requiring reason* in φ's favour. A requiring reason is, in short, a reason which speaks in favour of φ being required: in the absence of other considerations, such a reason will generate a requirement that I φ.

It might also prove true, of course, that φ is highly personally costly to me. Perhaps φ would cost me my left pinkie, for instance. If this is right, there could be a *cost-based permitting reason* in favour of choosing the alternative act ψ. A cost-based permitting reason is a reason which, in virtue of prudential costs, helps generate a permission in favour of an act which would have otherwise been forbidden. Thus, a cost-based

[12] Pummer (2023, 2025), who labels the general framework the *Feature Choice View*. Pummer's account, and the discussion of following sections, is developed within the framework of requiring versus permitting reasons. On the general distinction between the requiring versus permitting force of practical reasons, see Gert (2003).

permitting reason could generate a permission in favour of ψ even when the weight of requiring reasons otherwise speaks in favour of φ.

If there only existed requiring reasons and cost-based permitting reasons then you would indeed, it seems, be required to become a doctor in **Magazine**. The requiring reasons would pull towards doctoring, and no cost-based permitting reasons would speak against such a requirement. Pummer follows other moral autonomy theorists, however, in taking seriously the commonsense judgement that it would be permissible to pursue a career other than doctoring. In order to accommodate this judgement in **Magazine** and related cases, Pummer introduces a further species of permitting reason. These are what he labels *autonomy-based permitting reasons.*[13]

An autonomy-based permitting reason is a reason you have in favour of performing some act φ (rather than ψ) precisely in virtue of the fact that φ would garner a *distinctively valuable feature* for your future life, where such a feature would not have obtained under the alternative act ψ. Autonomy-based permitting reasons of this kind can generate a permission that you φ, even when the balance of requiring reasons would have otherwise required that you ψ. The label 'autonomy-based permitting reason' is apt since the core idea is that moral agents should be afforded, to at least some degree, a certain kind of creative moral freedom when it comes to determining what sorts of distinctive valuable features their future lives should contain.

If we return to **Magazine** with the tool of autonomy-based permitting reasons in our belt, we can justify and explain the commonsense verdict that both flute-playing and doctoring are permissible. The career of flute-playing contains countless distinctive features. There are thus very strong autonomy-based permitting reasons in its favour.

[13] The most fine-grained details are provided in Pummer (2025:198–202).

These autonomy-based permitting reasons are likely to undercut the requiring reasons which pull towards doctoring. In this way, **Magazine** is likely to contain a permission to engage in flute-playing rather than in doctoring – as well, of course, as a permission to engage in doctoring rather than flute-playing.

Two brief points about the above framework. First, importantly, autonomy-based permitting reasons cannot have a cancelling structure: they cannot simply cancel one another out when some such reasons pull towards φ and others pull towards ψ. Why? Well, in **Magazine**, there are autonomy-based permitting reasons in favour of becoming a doctor *and* in favour of becoming a flautist. After all, both lives contain countless distinctively valuable features. So if autonomy-based permitting reasons had a cancelling structure, these respective autonomy-based reasons would cancel one another out, and we would be left again with a determinate requirement to pursue doctoring. Since this is the result which moral autonomy theorists wish to avoid, autonomy-based permitting reasons cannot have a cancelling structure.

Second, Pummer proposes certain potential addendums to the above account of autonomy-based permitting reasons.[14] One addendum – call it the *Similarity Addendum* – says that the strength of my autonomy-based permitting reasons in favour of φ depends upon the extent to which the valuable feature x secured by φ is similar to the alternative valuable feature y secured by ψ. Plausibly, if x consists of earthenware pottery and y consists of stoneware pottery, then I do not have substantive autonomy-based reasons in favour of performing φ rather than performing ψ, where this result is explained precisely by the fact that the objective differences between earthenware pottery and stoneware pottery are small.[15] The second addendum – call it the *Attachment Addendum* – says that

[14] Pummer (2025: 199).
[15] This addendum raises deep issues about the measurement of similarity; I return to these issues in §5.

the strength of my autonomy-based permitting reasons in favour of φ are a function of my level of attachment to the valuable feature x secured by φ. If I am not currently attached to the valuable features secured by φ, then this weakens my autonomy-based permitting reasons in favour of performing φ. I am unsure, in the end, whether we should endorse the Attachment Addendum. The Attachment Addendum seems to imply that those who currently feel flute-neutral might lack a moral permission to play the flute, and this is a result about which we might feel ambivalent. Nonetheless, the Similarity Addendum seems highly intuitively plausible: furthermore, it will prove crucial in following discussion, raising its head again shortly.

3. Generating Lesser Numbers Permissions

As foreshadowed, my central purpose in this discussion is to show how moral autonomy theorists are committed towards LNPs across a wide spectrum of practically relevant rescue cases.

The argument comes in two steps. The first step is to show that, in canonical rescue cases like **Island**, there are indeed autonomy-based permitting reasons in favour of saving the few. The second step is to show that, in canonical rescue cases like **Island**, the autonomy-based permitting reasons in favour of saving the few must be strong enough to undercut the requiring reasons in favour of saving the many. If this is right, then moral autonomy theorists are indeed committed to affirming the widespread existence of LNPs.

3.1 Autonomy-Based Permitting Reasons for Saving the Few

The first step: to show that there are indeed autonomy-based permitting reasons for saving the few in canonical rescue cases like **Island**. Recall that, in **Island**, your choice is to save Alice from one island or Blaise, Charlie, Daniella, Ethan, and Francis from the

other. For colour, let's fill in the details a little further. Let's suppose that you know a fair bit about each of the persons who stand to be affected by your choice. Let's suppose, too, that you'll have a chance to meet and interact with whichever person you decide to save, after the fact.

Here is something which certainly seems true. If you decide to save Alice rather than the others, your life will contain distinctively valuable features which it otherwise would have lacked. What are these features? The most obvious feature is a particular kind of accomplishment: the accomplishment of *rescuing Alice*. To achieve this accomplishment is, *ipso facto*, to achieve certain other kinds of accomplishment. For instance, to rescue Alice is to spare Alice's parents from grief. It is also to preserve the many valuable long-term relationships which hold between Alice and other persons.[16]

Plausibly, though, the list of distinctive valuable features involved in saving Alice need not include only accomplishments. Most notably, the list may include certain kinds of morally important experiences. It may include, for instance, the experience of *meeting Alice*. Perhaps it will include the experience of seeing Alice reunited with her family, or of eventually hearing about the many wonderful things that Alice has gone on to do with her life, now that she has a life to live in the first place.

If this is right, though, then there are going to be autonomy-based permitting reasons in favour of saving Alice. After all, according to moral autonomy theorists, the presence of distinctive valuable features such as these *just is* what generates autonomy-based permitting reasons in favour of performing one act rather than another. So long as there are distinctive valuable features involved in saving just Alice, there are going to be

[16] Throughout, I follow Pummer (2025: 198) in using the word 'feature.' In other contexts, the word 'feature' seems to pick out something closer to 'property.' For current purposes, it picks out something closer to 'particulars.' Whichever label we use, the category must presumably be broad enough to include more than just chairs and tables: as Pummer notes, it must be able to include things such as activities, events, states of affairs, accomplishments, and so on.

autonomy-based permitting reasons in favour of saving just Alice. I haven't yet said anything about the strength of these autonomy-based permitting reasons. Perhaps these reasons are exceedingly weak. But it seems right, as a starting point, to acknowledge that such reasons exist.

Of course, cases like **Island** are rare and unusual things. It's virtually never the case that we have the chance to costlessly save the one rather than the five *via* direct rescue. By contrast, each of us is confronted with *indirect* opportunities to costlessly save the one rather than five on a regular basis. Thus consider:

> **Donation.** You are deciding whether or not to become a regular donor for famine relief versus pandemic prevention programmes. Your giving would be made on a weekly basis over the coming year. Donating to only one charity won't make you strictly worse-off since, within reason, regular charitable donation tends to be associated with well-noted welfarist benefits. Donations towards famine relief would save one life (in expectation) and donations towards pandemic prevention programmes would save five lives (in expectation).[17]
>
> **Volunteering**. You are deciding whether to volunteer as a fundraiser for famine relief versus pandemic prevention programmes over a period of one year. Volunteering will take several hours each week. Volunteering for just one cause won't make you strictly worse-off since, within reason, regular volunteering turns out to be associated with well-noted welfarist benefits. Volunteering for famine relief would save one life (in expectation) and volunteering for pandemic prevention would save five lives (in expectation).[18]

In **Donation** and **Volunteering**, you are deliberating about where to gift some marginal portion of your income and labour. You will be just as well off wherever you donate your income and labour. But nonetheless, donating your income and labour towards different causes will save different numbers of lives (in expectation).

---

[17] For a survey on the welfarist benefits of regular giving, see MacAskill, Mogensen & Ord (2018). We can assume that you don't have the money to donate to both charities, or alternatively, that donating to both charities would begin to generate welfarist costs such that there is a sufficiently weighty cost-based permitting reason against donating. I only pick famine relief and pandemic prevention as token examples, without meaning to imply anything concrete about the comparative cost-effectiveness of these charitable avenues in the real world. We are simply in want, here, of different lifesaving charities for which there is an asymmetry in cost effectiveness.

[18] For some of the literature on the welfarist benefits of prosocial behaviour, see Chapter 1.

The key point I want to make here is that, plausibly, just as you would have autonomy-based permitting reasons for costlessly saving the few in **Island**, you would have autonomy-based permitting reasons for costlessly saving the few (in expectation) in **Donation** and **Volunteering**. Indeed, if anything, it seems that your autonomy-based permitting reasons for saving the few would be *stronger* in these practically-relevant case types.

Why? Well, upon reflection, it's not hard to see how regular donation and volunteering for (say) famine relief *in particular* might secure many distinctively valuable features for your future life. To engage in regular donation and volunteer work for famine relief involves achieving certain distinctive accomplishments: it is to make a positive causal difference, for instance, in the supply chains of famine relief charities. To the extent that regular donation and volunteering for some cause *x* causes you to become more informed about *x*, then donating and volunteering for famine relief will also allow you to gain various kinds of important knowledge. (You will learn, perhaps, about the nature of famine, the ways in which famine relief charities work, the structure of agricultural supply chains, and so on.) One possibility is that regular donation and volunteering for famine would help you in cultivating certain noble dispositions – a disposition to waste less, for instance. And another possibility is that regular donation and volunteering for famine relief could play a role in your forming a certain kind of social identity: your identification as someone who cares *about famine*. All of these, it seems, are distinctively valuable features which would accrue for your future life if you gave your time and money to famine relief in **Donation** and **Volunteering**.

If this is right, though, then there are going to be autonomy-based permitting reasons in favour of donating and volunteering for famine in both **Donation** and **Volunteering**. After all, according to moral autonomy theorists, the presence of distinctive features such

as these *just is* what generates autonomy-based permitting reasons in favour of performing one act rather than another. Again, I haven't yet said anything at all about the strength of these autonomy-based permitting reasons. Perhaps they are exceedingly weak. But it seems right, as a starting point, to acknowledge that such reasons exist.

3.2 Permissions to Save the Few

I've shown how, under the above framework, we would systematically possess autonomy-based permitting reasons for rescuing the few rather than the many. The second step of the argument, recall, is to show that these autonomy-based permitting reasons must be strong enough to undercut your requiring reasons in favour of saving the five. This may sound like a tall order. But it turns out to be doable. The crucial trick, here, is to draw upon some of the independent judgements which moral autonomy theorists were already inclined to make.

Here is how the reasoning works. I began the discussion, recall, by noting that many moral autonomy theorists have a fairly liberal view of permissibility when it comes to grand and sweeping career choices like **Magazine**. As I put the point previously: moral autonomy theorists tend to assert that, when choosing between different (non-harmful) career paths, any choice made ends up being innocent. Let's thus assume, in what follows, that moral autonomy theorists are committed towards the following claim: in the actual world, it is in fact permissible for me to pursue a career of flute-playing. This is true even if, in the actual world, various alternative lifesaving careers (such as doctoring, charity fundraising, and so on) turn out to be just as prudentially good in expectation.[19]

[19] For instance, in his original defence of moral autonomy, Slote (1985: 26–27) judged that *any* non-harmful career path is morally permissible, and extended this claim to even include those careers which are prudentially worse for the agent by some degree. I am not attributing to moral autonomy theorists a claim this strong, but only the weaker claim that equally prudentially good (non-lifesaving) careers are to be viewed as permissible in the actual world. A similar judgement is reported in Pummer (2025: 187–188). Lazar (2019: 83) claims to introspect that it was not wrong for him to have sought out the career of (non-

What does this judgement imply? Well, it implies that your autonomy-based permitting reasons in favour of pursuing "career flute-playing" must be *exceedingly strong*. For it's quite likely, upon reflection, that other career paths within your reach could have saved very many lives in expectation. Consider the following toy example. Suppose that one of your (equally prudentially good) career options is to work as a professional fundraiser for anti-malarial charities on an average lifetime annual salary of £40,000. The average return on investment (ROI) for charities in the United Kingdom lies in the region of 4:1 – that is, for every £1 spent on fundraising, charities typically receive £4 in returns.[20] In extremely loose terms this suggests that, minus your salary, your professional labour might still garner in excess of £100,000 for anti-malarial charities for each year in which you work. Even assuming that you only fundraise for thirty professional years – you enter the career a little late and require training, say – this suggests that your total lifetime labour could raise in the region of £3,000,000 in funds for anti-malarial charities. This amounts to very many saved lives. The most reputable charity evaluators typically estimate, for instance, that it costs somewhere in the region of £3000 to save each additional marginal life in expectation via the most effective forms of malarial intervention.[21] Under these numbers, then, your lifetime professional labour might end up saving somewhere in the region of one thousand lives in expectation.[22]

---

lifesaving) philosophy: I suspect that Lazar would wish to maintain as much even if certain alternative (lifesaving) careers turned out to be just as prudentially good as philosophy, since Lazar is sympathetic, on the next page, to the category of self-sacrificing permissions.

[20] See, for instance, the 2024 UK Civil Society Almanac: https://www.ncvo.org.uk/news-and-insights/news-index/uk-civil-society-almanac-2024/. Of course, the 4:1 figure masks numerous complexities beyond the scope of current discussion.

[21] See, for instance, https://www.givewell.org/charities/top-charities.

[22] There are many difficult complications here. One such complication is that, for many altruistic jobs, had you not pursued that job, somebody else would have done so. This would seem to suggest that the counterfactual difference made by your labour is quite small. On the other hand, though, this complication begets further complications. Insofar as someone else ends up having to pursue the job that *you* didn't pursue, this prevents them from pursuing a *different* altruistically-focussed career somewhere else, and so on. This seems to suggest that the counterfactual difference made by your labour is still very large. It's difficult to know what to say, here, beyond the observation that there are certain ways in which you might

Of course, assigning precise numbers to these things is notoriously difficult. It inevitably introduces artificiality into the discussion. The point here isn't to get bogged down in the numbers, but merely to provide a very rough sense of just how strong autonomy-based permitting reasons are going to have to be if they are going to render non-lifesaving careers permissible in grand choices like **Magazine**. The distinctive features involved in thirty years of "career fluting", going by the above numbers, may have to generate autonomy-based permitting reasons strong enough to undercut the requiring reasons in favour of saving somewhere in the region of a thousand lives (in expectation). This suggests that, on average, the distinctive features involved in "career fluting" for *one* year may have to generate autonomy-based permitting reasons strong enough to undercut the requiring reasons in favour of saving *thirty-three* lives (in expectation).

If these are the sorts of weightings to which moral autonomy theorists are committed, though, then we should indeed expect very substantial LNPs to emerge in cases like **Island**, **Donation**, and **Volunteering**. Consider **Volunteering**, for instance. What we've already seen is that volunteering for famine relief gives your life a set of highly distinctive features over the period of one year. Perhaps these features don't hold the same weight as (say) the features involved in pursuing "career fluting" for a year. But it seems plausible to me that they might reasonably hold (say) *one quarter* of the weight of the features involved in pursuing "career fluting" for a year. If this is right, then the autonomy-based permitting reasons in favour of volunteering for the less effective charity, in **Volunteering**, would still be more than strong enough to outweigh the requiring reasons

labour as a career altruist which seemingly aren't subject to overdetermination: working within under-resourced altruistic start-ups, for instance, or helping to build altruistic initiatives from scratch.

in favour of saving eight additional lives. Moral autonomy theorists, in such a case, would be committed to the presence of very substantial LNPs.

So much for **Volunteering**. But what about **Island** and **Donation**? These cases are a little harder to evaluate. Nonetheless, it still strikes me that autonomy-based permitting reasons might generate substantial LNPs within these cases. Consider **Donation**. I've already noted that becoming a regular and conscientious donor for famine relief could bring certain distinctively valuable features into your future life. It could involve achieving certain distinctive kinds of accomplishments, involve gaining access to certain distinctive kinds of knowledge, forming certain distinctive noble dispositions, and building a distinctive kind of social identity. It's unlikely, I think, that the distinctive features garnered by regularly donating for famine relief could be as weighty as those various distinctive features garnered by *actively volunteering* for famine relief (for the purposes of generating autonomy-based permitting reasons). But note that these features wouldn't need to be that strong in order to generate LNPs. For instance, suppose that the distinctive features garnered by regularly donating towards famine are only *half* as weighty as the distinctive features garnered by actively volunteering for famine (for the purposes of generating autonomy-based permitting reasons). If this is right, then given the numbers within the previous paragraph, your autonomy-based permitting reasons for donating to famine in **Donation** would still prove weighty enough to undercut your requiring reasons in favour of saving four additional lives.[23]

[23] This seems to imply that Pummer (2025: 202) cannot quite claim, at least without further argument, that autonomy-based permitting reasons are generally compatible with a conditional requirement to donate marginal funds in the more cost-effective manner. This result would only hold if we could show, in general, that donating $\$n$ towards one lifesaving charity *A* brings the same set of valuable features into your life as does donating $\$n$ towards a different lifesaving charity *B*. For the reasons I've begun to sketch, this claim does not strike me as persuasive.

This just leaves **Island**. Overall, I think it is very difficult to evaluate the strength of your autonomy-based permitting reasons in favour of saving the few in **Island**. Why? Well, on the one hand, the emergency in **Island** may be over in a matter of minutes. If this is right, it is tempting to claim that the 'distinctive valuable features' secured by saving Alice must be featherweight ones. On the other hand, there is no shortage of people who report that directly saving a life is one of the most impactful and profound experiences they have ever undergone. If this is right, it is tempting to claim that the 'distinctive valuable features' secured by saving Alice must be heavyweight ones.[24]

Is there a way of making progress on the puzzle? One possible heuristic for finding the 'weight' of any given distinctive feature x, as grounds for autonomy-based permitting reasons, would be to ask how often and how intensely x would serve as the basis of pro-attitudes (admiration, loyalty, respect, and so on) within the agent in question, were she to gain access to that feature. For instance, suppose that I am choosing between a pet dog for three years versus a pet cat for one year. The fact that a pet dog would ground pro-attitudes on my part for *longer* would imply that the pet dog is a more weighty distinctive feature within my future life, at least for the purposes of generating autonomy-based permitting reasons. Similarly, if I am disposed to feel more intense admiration for dogs than for cats, then this too would imply that the pet dog is a more weighty distinctive feature within my future life, at least for the purposes of generating autonomy-based permitting reasons.

If this picture is broadly right, though, then it is plausible that you have fairly weighty autonomy-based permitting reasons for saving just Alice in **Island**. After all, perhaps the accomplishment of saving the particular person Alice will only 'occupy your mind' for a

[24] For those who feel that the 'distinctive valuable feature' of saving Alice's life in **Island** is basically the same 'distinctive valuable feature' as saving the lives of the many: hold fire a moment longer, till §4.

period of two months; after this, the incident will lose all practical relevance within your daily life. Nonetheless, plausibly, the pro-attitudes sparked by the accomplishment of directly saving Alice will be of a much more intense sort than (say) the pro-attitudes sparked by the making of regular charitable donations. It is, after all, an accomplishment which will viscerally and powerfully shape the character of your life *across* that short window of time:

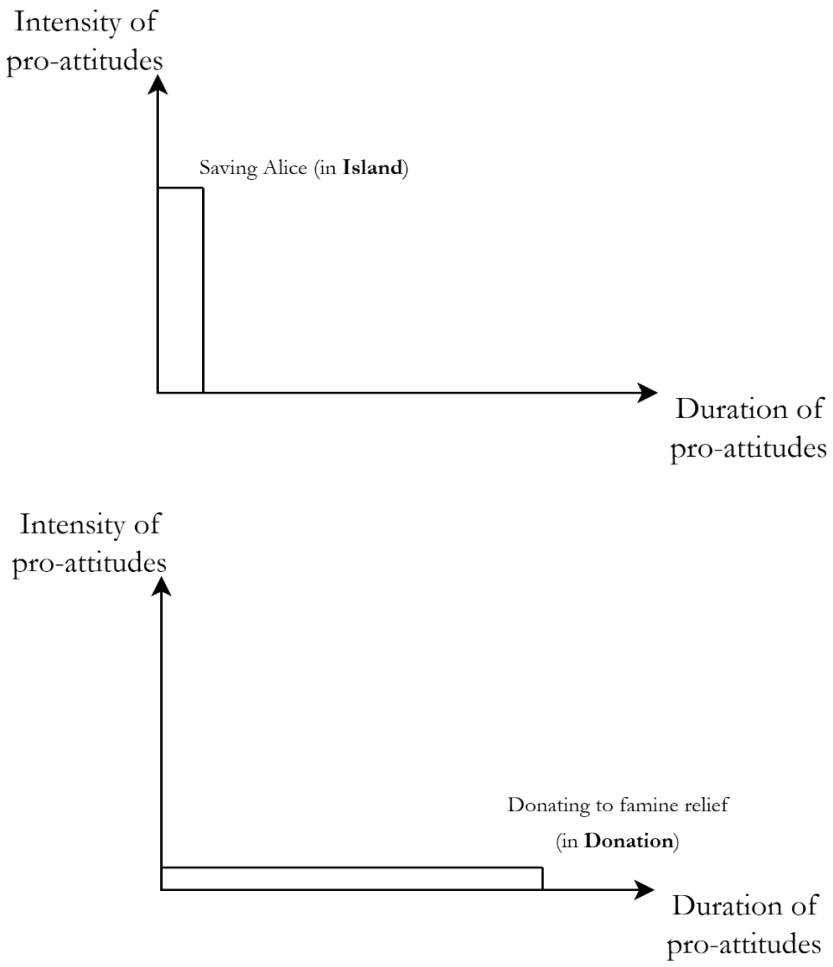


fig. 1. Intensity and duration of pro-attitudes caused by valuable features, in ***Island*** *and* ***Donation.***

My own sense, then, is that the distinctive valuable features involved in directly saving a life in **Island** should be *at least* as weighty, for the purpose of generating autonomy-based permitting reasons, as the distinctive valuable features involved in indirectly saving lives over the course of a year in **Donation**. Although the former features might generate pro-attitudes which peter out more quickly, they generate pro-attitudes of a much more intense kind. If this is right, then your autonomy-based permitting reasons in **Island** would indeed be strong enough to outweigh the requiring reasons in favour of saving four additional others – as a minimum.

4. The Similarity Objection

I've shown how autonomy-based permitting reasons would systematically generate LNPs: that is, systematically generate permissions to costlessly rescue the few rather than the many. This result hasn't just held in **Island**; it has also held in more practically relevant rescue cases such as **Donation** and **Volunteering**.

So far, I've skated over an important objection. The objection has the following shape: "Suppose you save the lesser number in **Island** or **Donation** or **Volunteering**. Perhaps this will indeed cause your future life to contain certain valuable features. But nonetheless, these features will be *very similar* to the features secured by saving the greater number. Since the features secured by saving the lesser number are so similar to the features secured by saving the greater number, any autonomy-based permitting reasons in favour of saving the lesser number ought to be heavily discounted. These (heavily discounted) autonomy-based permitting reasons will be incapable of generating a permission to save the lesser number."

Call this the *Similarity Objection*. The objection is worth taking seriously. The objection gains traction because, as we already saw in Pummer's original articulation of autonomy-

based permitting reasons, it seems that considerations of similarity *should* lead us to discount autonomy-based permitting reasons. If some act φ secures life feature *x* (earthenware pottery) and the alternative act ψ would have secured a highly similar life feature *y* (stoneware pottery), then it seems that the autonomy-based reasons generated by *x* and *y* ought to be discounted, to at least some degree.[25]

I don't quite find the Similarity Objection persuasive. Here is why. First, let's start by assuming that the similarity discount works in the following way: we must discount some life feature *x* (garnered by φ) in comparison with life feature *y* (garnered by ψ) as grounds for autonomy-based permitting reasons insofar as my *qualitative experience* of having feature *x* would be similar to my *qualitative experience* of having feature *y*.[26]

If this is how the objection works, then I don't think the Similarity Objection would undermine the claims that I have made concerning (for instance) **Volunteering** and **Donation**. I have already suggested, for instance, that the qualitative experience of volunteering for famine relief in **Volunteering** is likely to be highly different from the qualitative experience of volunteering for pandemic prevention programmes. Volunteering for famine relief, for instance, will involve gaining knowledge about distinctively valuable domains, having the chance to actively engage in benefitting

[25] One relevant complication for any proponent of the Similarity Objection – and indeed, the Similarity Addendum itself – is that our evaluations of similarity are highly sensitive to the context in which they are being made. In general terms, resemblists are happy to grant that there is an objective, context-independent fact of the matter about the level of similarity which holds between any given pair of objects, where this fact of the matter is determined by considerations such as (say) the number of sparse or natural properties held in common by those objects. For discussion, see Blumson (2018). Anti-resemblists, following Goodman (1972: 445), tend to deny that there is an objective, context-independent fact of the matter about the level of similarity which holds between objects.

[26] Pummer's (2025: 199) suggestion, for instance, is that what is relevant in measuring 'difference' between life features is the variance in both their qualitative and durational aspects. One could, perhaps, precisify Pummer's suggestion about qualitative difference in two different ways: either what matters is the difference in one's qualitative *experience* of the good itself (earthenware pottery and stoneware pottery are no different if I cannot tell the difference) or what matters is qualitative difference between *the objects themselves* (earthenware pottery and stoneware pottery still count as 'different' even if I cannot tell the difference). In what follows, I consider both kinds of suggestion.

particular persons within particular geographic regions, having the chance to refine distinctive sets of practical skills, having the chance to form a distinctive social identity, and so on. For these reasons it would not be plausible, I think, to try and apply any substantive similarity discount to the life features on offer in **Volunteering**.

To some extent, I think that similar points could be made concerning **Donation**: to some extent, I think the qualitative experience of regularly and conscientiously donating towards famine is likely to be qualitatively different from the qualitative experience of donating towards a different charitable cause. As we've already seen, to miss out on becoming a regular and conscientious giver towards (say) famine relief might be to miss out on the feeling that one has helped achieve certain distinctive accomplishments, to miss out on the chance to access to certain distinctive kinds of knowledge, … and so on. But importantly, even if we grant that *some* level of similarity discounting is appropriate in **Donation**, as may be the case, this fact alone wouldn't undermine the argument for LNPs. For instance, I previously concluded that your autonomy-based permitting reasons in favour of donating for the few in **Donation** must, given the independent judgements made by moral autonomy theorists, be strong enough to outweigh the requiring reasons in favour of saving four additional others. If we were thus to *halve* the weight of these autonomy-based permitting reasons in light of similarity considerations, **Donation** would still end up containing a permission to refrain from saving two additional others. We would still arrive at an LNP, though of a weaker kind than before.

But second, and perhaps more importantly, it is not obvious to me that the similarity discount *should* track qualitative experience. It's not obvious to me, as I put things a moment ago, that we should: "discount some life feature x (garnered by φ) in comparison with life feature y (garnered by ψ) as grounds for autonomy-based permitting reasons

insofar as my *qualitative experience* of having x would be similar to my *qualitative experience* of having y."

It often seems, for instance, that the similarity discount should instead track something much closer to the *objective* level of difference between features x and y themselves, as opposed to my qualitative experience of having access to either feature.[27] To see as much, consider a deathbed case in which a dying patient can press Button A or Button B. The qualitative experience of pressing each button is highly similar – both buttons are blue, round, and made of plastic, for instance. Nonetheless, pressing one button will substantially advance famine relief programmes. Pressing the other button will substantially advance pandemic prevention programmes. The proponent of moral autonomy, I take it, will want to maintain that there are strong autonomy-based permissions to press either button, even though there is a natural sense in which the qualitative *experience* of pressing each button is rightly viewed as being highly similar.[28]

If this is right, then I think it is a stretch to try and apply a substantial similarity discount to cases like **Donation** and **Volunteering** – indeed, even to apply a substantial similarity discount to easy rescue cases like **Island**. For instance, at the moment of choice in **Island**, the qualitative experience of dragging one person aboard the port side might be reasonably similar to the qualitative experience of dragging four other persons aboard the starboard side. Nonetheless, it would not follow that it is appropriate to apply a similarity discount to those valuable features ('saved lives') secured by either act. That would only be true if, in the relevant objective sense, *persons themselves* were appropriately

[27] Again, see fn.25 on some of the complexities in speaking of 'objective' similarity.
[28] Perhaps the experience of pressing each button is in fact highly qualitatively different because one *understands* each button press to represent the realisation of different moral goods. There is certainly nothing wrong with this suggestion. But importantly, it would imply that the qualitative experience of (say) donating to the less effective charity in **Donation** is highly different to the qualitative experience of donating to the more effective one, even if donating only takes mere seconds. One *understands*, in such a case, that each donation represents different the preservation of different moral goods.

characterised as 'highly similar.' This is quite a hard thing to believe. It is a very familiar suggestion that persons, and the value contained within their separate lives, are *non-fungible*. Unlike porcelain cups, you cannot break a human, throw it out, and replace it with another without having suffered a loss.[29] One natural upshot of humans being non-fungible, I suspect, is that it is ethically inappropriate to try and discount human patients as 'highly similar' *qua* objects of value when we deliberate about the different possible outcomes which might follow our acts. It is a very different thing – namely, a different *person* – who is saved if I sail towards the island containing the few rather than the many.

5. Existing Arguments

What I've shown, then, is that moral autonomy theorists are systematically committed to belief in LNPs. For want of a label, let's call the above line of reasoning the *Autonomy Argument* in favour of LNPs. Of course, the Autonomy Argument would hardly be the first argument in favour of LNPs. Before closing, it may be particularly helpful to draw out some of the important ways in which the Autonomy Argument differs from *existing* pro-LNP arguments.

A natural starting point, here, is the detailed set of arguments for LNPs developed in landmark work by John Taurek.[30] It's difficult, admittedly, to tease apart the different threads of Taurek's original discussion. One Taurekian argument for LNPs would claim, roughly speaking, that the strength of your requiring reasons in favour of saving the five must be exceedingly weak. For it's a common intuition that it would be permissible to save Alice insofar as she turns out to be a mere acquaintance of yours. But if this is right it means that your requiring reasons in favour of saving the five, if there are such things,

[29] For relevant discussion, see Hare (2016), and recently, Tucker (n.d.).
[30] Taurek (1977).

can be overridden by even the most exceedingly "feeble" kind of social relations (the *Feeble Relations Argument*).[31] Another Taurekian argument observes that there is nobody who is benefited *more* if you save the five, since there is no such thing as the aggregative perspective from which benefits and losses can be measured. In light of this fact it's just as choiceworthy to save the one as it is to save the five: perhaps one ought to navigate decisions like **Island** by flipping a coin (the *Anti-Aggregative Argument*).[32]

Perhaps there is some relation between the above Autonomy Argument and Taurek's Feeble Relations argument: Taurek's Feeble Relations argument may partly be getting at the idea that our requiring reasons for saving the many can never be strong enough to outweigh our autonomy-based permissions to choose *whom* we would like to save, including mere acquaintances or complete strangers. But the most important point to make here is that the Autonomy Argument for LNPs gains traction entirely independently of the Anti-Aggregative Argument. In particular, the Autonomy Argument gains traction even insofar as one *does* believe in aggregation, assuming, as my discussion has so far done, that your requiring reasons for saving the five are much stronger than your requiring reasons for saving the one. In precisely such an aggregative context, autonomy-based permitting reasons have turned out to generate sizable LNPs.

Of course, staunch Taurekians may not feel much interest in the Autonomy Argument, since they may already take their Anti-Aggregative Argument to generate LNPs across the various case types in which I've been interested – across cases like **Island**, **Donation**, and **Volunteering**. I don't think, however, that this shows the Autonomy Argument to be unimportant. This is for the simple reason that many, including myself, do not feel the

[31] Although Taurek (1977: 298) doesn't employ the language of requiring versus permitting reasons; I take it that this is one way of interpreting the suggestion.
[32] For recent discussion of whether Taurek does in fact endorse coin flipping, see Sung (2022).

pull of full-blooded Taurekian anti-aggregationism.[33] Those who reject full-blooded Taurekian anti-aggregationism are still left, it seems, to grapple with the challenge posed by the Autonomy Argument.

Of course, since Taurek's original discussion, various other arguments have been developed in favour of LNPs.[34] Perhaps the most influential recent argument in favour of LNPs is developed in the work of Kieran Setiya.[35] What's particularly worth drawing out, here, is how the Autonomy Argument would vindicate a *much larger* set of LNPs than does Setiya's argument. Let me explain.

Setiya's argument for LNPs – call it the *Love Argument* – begins with some observations about the characteristic nature of love. For Setiya, love consists of an arresting recognition of the value of another's humanity. It is sufficient to justify and account for my love for you simply that I see you and find myself struck by your value as a precious (rational) creature.[36] In this way, love can be characterised as having an unconditional nature: love does not require that the beloved be sufficiently well-known, nor does it require that the beloved possess some sufficiently desirable set of qualities.

This kind of love, suggests Setiya, can make it rational and reasonable for me to display a special concern towards your well-being. This can be the case even in the absence of any prior relationship.[37] Of course, prior special relationships such as marriage or parenthood might *amplify* our moral expectations when it comes to special treatment. But nonetheless, even in the absence of such prior relationships, the presence of love renders it rational or reasonable to show a special concern for the well-being of another person.

[33] Of many recent examples, see Barnett (2025).
[34] See fn.4.
[35] Setiya (2014).
[36] Setiya (2014: 260–263).
[37] Setiya (2014: 263–270). C.f. Kolodny (2003).

Setiya's liberal portrait of the justifying conditions of love, it turns out, often implies LNPs. Take **Island**, for instance. By assumption, the potential beneficiaries in **Island** are each perfect strangers. Nevertheless, even if Alice is a perfect stranger, I may see her and feel unconditional love towards her. This love could emerge even in the absence of a prior relationship. Love of such a kind would be sufficient to make reasonable and rational a display of special concern for Alice's welfare – it would make it reasonable and rational to save *Alice* rather than the five, in much the same way that it might be reasonable and rational to save my beloved friend rather than the five.

The crucial point to emphasise, here, is that there are many practically relevant cases in which the Love Argument appears incapable of generating LNPs but in which the Autonomy Argument nevertheless *does* seem to generate LNPs. This is because Setiya's argument only seems to generate LNPs within cases where agents can form love attitudes towards those they aid. The Autonomy Argument, by contrast, is governed by no such attitudinal constraint.

It's a tempting thought, for instance, that I can only form love attitudes towards beneficiaries so long as I can think *singularly* about those beneficiaries, where singular thoughts, roughly speaking, are those thoughts *directly about* a given referent, facilitated by some process of acquaintance, perceptual or otherwise.[38] It is this fact which explains why I can feel love towards a neighbour with whom I have regular face to face interactions, but why I cannot feel love towards "the first stranger to be born at the turn of the next century."[39] But if this is right, the Love Argument will struggle to generate LNPs in some of those practically-relevant case types in which I have been most

[38] For an overview, see Jeshion (2010: 105–140).

[39] Of course, we should be careful not to overstate the point: there are many other kinds of concern attitude which we might hold towards beneficiaries about whom we cannot think singularly. I might care a great deal, for instance, about preventing the assassination of the stranger born next century. For now, the suggestion is simply that there is a distinctive kind of love attitude (capable of grounding partial preference) mediated via direct acquaintance. Recently, see Setiya (2023).

interested: cases such as **Donation** and **Volunteering**. After all, insofar as cases like **Donation** and **Volunteering** are glossed as involving the provision of indirect international aid, then the cases will normally preclude acquaintance with beneficiaries at the moment of choice. Such cases will thus preclude singular thought about beneficiaries at the moment of choice. They will thus preclude love attitudes towards beneficiaries at the moment of choice – or at least, the sorts of love attitudes which appear to be operative in Setiya's argument. The Love Argument will hence fail to generate LNPs in **Donation** and **Volunteering**.

By contrast, **Donation** and **Volunteering** were cases in which the Autonomy Argument was on firm ground: these were cases in which autonomy-based permitting reasons *did* seem capable of undercutting the requiring reasons in favour of saving five others. For these were cases in which saving the few garnered distinctively valuable features for the course of your future life. Features such as these, it seems, can be present regardless of whether or not you are capable of forming love attitudes towards the particular beneficiaries whose interests are at stake in your choice. They can come about even if one aids the unknown from a very great distance.

6. Conclusion

In the above discussion I've shown that moral autonomy theorists must systematically accept LNPs: they must systematically accept that it is permissible to rescue the one rather than the five. This result hasn't just held in highly idealised rescue cases like **Island**. It has also held in practically relevant rescue cases such as **Volunteering** and **Donation**.

As foreshadowed at the outset, I am inclined to view this as something of a dilemma. On the one hand, I am occasionally inclined to join with Pummer and others in

emphasising the importance of moral autonomy.[40] On the other hand, like many philosophers, I want to uphold a requirement to save the greater number in easy rescues like **Island**. I've ultimately suggested that this tightrope is unwalkable: we must either reject the appeal to moral autonomy in cases like **Magazine** or we must accept that one may save the lesser number in easy rescues such as **Island**.

I am tentatively inclined to resolve the dilemma by rejecting the appeal to moral autonomy in cases like **Magazine** – more generally, by rejecting the view that considerations of moral autonomy could be of substantive moral importance. This stems from my own sense that LNPs are simply very difficult to accept. But it is of course impossible to stop the moral autonomy theorist from taking things in the other direction: from concluding that it must indeed be generally be permissible to save the lesser number, and that the case for LNPs is on much stronger ground than has been historically appreciated.

As a final note, let me briefly emphasise that the above discussion is clearly going to have implications for recent debates concerning the effective altruist movement.[41] Though definitions are difficult, put very broadly, effective altruism constitutes an international project in which individuals, using the best available empirical tools and methods as their guide, attempt to use their resources to do as much good as possible.

One potential interpretation of the above discussion is that moral autonomy theorists cannot be effective altruists. The reasoning would go as follows: "Effective altruists maintain that we are morally required to costlessly save the greater number. But as we have seen, moral autonomy theorists systematically maintain that it is permissible to

[40] In my Chapter 1 I make the important observation that agents in the real world do not face idealised binary choices like **Magazine**: they often seem to be faced with a range of prudentially good, qualitatively diverse, altruistically-focussed lifestyles. In those real-world cases, I feel more comfortable with a requirement that one must restrict one's attention towards *an* altruistically-focussed lifestyle, and find that the appeal to moral autonomy less persuasive.

[41] For an overview, see Greaves & Pummer (2019).

costlessly save the lesser number. Ergo, one cannot be both a moral autonomy theorist and an effective altruist."

I am not entirely convinced by this interpretation. Many effective altruists, for instance, explicitly refrain from claiming that we are *morally required* to maximise the good. According to at least one informal survey, 37.7% of effective altruists explicitly deny that they view effective altruism (understood as the project of doing as much good as possible) as a kind of 'moral duty' at all, instead preferring to understand effective altruism as a kind of moral 'opportunity.'[42] Similarly, definitions provided by prominent effective altruists are often framed in explicitly non-normative terms, such that identifying as an effective altruist does not *entail* any particular view of our moral requirements.[43]

Suppose, as an empirical matter, that effective altruists were generally committed towards the following sorts of claims: it is *praiseworthy* to save the many in cases like **Island**, **Donation**, and **Volunteering**, in the sense that this behaviour ought to be pursued, encouraged, and promoted. Similarly: it is *blameworthy* to save the few in cases such as **Island**, **Donation**, and **Volunteering**, in the sense that (at least in comparison with saving the many) such behaviour ought to be avoided, discouraged, and criticised.

If these are the sorts of claim to which effective altruists are generally committed, then there may not be an inherent contradiction in being both a moral autonomy theorist and an effective altruist. The moral autonomy theorist could maintain, for instance, that we generally have a protected moral permission to save the lesser number. But they could also agree with the effective altruist that saving the greater number in such cases is praiseworthy and that saving the lesser number in such cases is blameworthy – perhaps in virtue of evincing a lack of concern for others, or in virtue of some other such

[42] McGeoch & Hurford (2017).
[43] As in MacAskill (2019).

consideration. In this way, the moral autonomy theorist might ultimately label acts which save the few as *suberogatory:* as acts which are permissible but nevertheless deserving of blame.[44]

Naturally, there are many potential choice points here. The important point to emphasise is that a contradiction only transparently emerges insofar as the effective altruist advocates for a much stronger claim: the claim that it is in fact *impermissible* to engage in non-maximally effective forms of charitable donation. If this is the claim for which the effective altruist is advocating – and I suspect it may often not be – then the above discussion would indeed show that moral autonomy theorists cannot be effective altruists.

[44] Driver (1992).

References

Anscombe, E. (1967) 'Who is Wronged? Philippa Foot and 'Double Effect'', *The Oxford Review*, *5:* 16–17.

Barnett, Z. (2025) 'Save the Five: Meeting Taurek's Challenge', *Philosophy and Phenomenological Research*, 110(2): 576–585.

Blumson, B. (2018) 'Two Conceptions of Similarity', *Philosophical Quarterly*, 68(270): 21–37.

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

Doggett, T. (2013) 'Saving the Few', *Noûs*, 47(2): 302–315.

Driver, J. (1992) 'The Suberogatory', *Australasian Journal of Philosophy*, 70(3): 286–295.

Foot, P. (1967) 'The Problem of Abortion and the Doctrine of Double Effect', *Oxford Review*, 5: 5–15.

Gert, J. (2003) 'Internalism and Different Kinds of Reasons', *The Philosophical Forum*, 34(1): 53–72.

Goodman, N. (1972) *Problems and Projects*. Bobbs-Merrill.

Greaves, H. & Pummer, T. (eds) (2019) *Effective Altruism: Philosophical Issues*. Oxford University Press.

Hare, C. (2016) 'Should We Wish Well to All?', *Philosophical Review*, 125(4): 451–472.

Henning, T. (2023) 'Numbers Without Aggregation', *Noûs*, 58(3), 755–777.

Hsieh, N., Strudler, A., & Wasserman, D. (2006) 'The Numbers Problem', *Philosophy & Public Affairs*, 34(4): 352–372.

Hurley, P. (1995) 'Getting Our Options Clear: A Closer Look at Agent-Centred Options', *Philosophical Studies*, 78(2): 163–188.

Jeshion, R. (2010) 'Singular Thought: Acquaintance, Semantic Instrumentalism, and Cognitivism', in R. Jeshion (ed), *New Essays on Singular Thought*. Oxford University Press.

Kagan, S. (1989) *The Limits of Morality*. Oxford University Press.

Kamm, F. (1985) 'Equal Treatment and Equal Chances', *Philosophy & Public Affairs*, 14(2): 177–194.

——— (1992) 'Non-Consequentialism, the Person as an End-in-Itself, and the Significance of Status', *Philosophy & Public Affairs*, 21(4): 354–389.

——— (1993) *Morality, Mortality: Vol 1: Death and Whom to Save From It*. Oxford University Press.

Kavka, G. (1979). 'The Numbers Should Count', *Philosophical Studies*, 36(3): 285–94.

Kolodny, N. (2003) 'Love as Valuing a Relationship', *Philosophical Review*, 112(2): 135–189.

Kumar, Rahul. (2001) 'Contractualism on Saving the Many', *Analysis*, 61(2): 165–170.

Lazar, S. (2019) 'Moral Status and Agent-Centred Options', *Utilitas*, 31(1): 83–105.

MacAskill, W. (2019) 'The Definition of Effective Altruism', in H. Greaves & T. Pummer (eds), *Effective Altruism: Philosophical Issues*. Oxford University Press.

MacAskill, W., Mogensen, A., & Ord, T. (2018) 'Giving Isn't Demanding', in P. Woodruff (ed), *The Ethics of Giving: Philosophers' Perspectives on Philanthropy*. Oxford University Press.

McGeoch, E., & Hurford, P. (2017). 'EA Survey 2017 Series: Distribution and Analysis Methodology,' *Effective Altruism Forum*, 29 August 2017. Available at http:// effective-altruism.com/ea/1e0/effective_altruism_survey_2017_distribution_and/.

Muñoz, D. (2022) 'The Many, the Few, and the Nature of Value', *Ergo: An Open Access Journal of Philosophy*, 9(4): 70–87.

Munoz-Dardé, V. (2005) 'The Distribution of Numbers and the Comprehensiveness of Reasons', *Proceedings of the Aristotelean Society*, 105: 191–217.
Otsuka, M. (2004) 'Skepticism About Saving the Greater Number', *Philosophy & Public Affairs*, 32(4): 413–426.
Parfit, D. (1978) 'Innumerate Ethics', *Philosophy & Public Affairs*, 7(4): 285–301.
Preston-Roedder, R. (2014) 'A Better World', *Philosophical Studies*, 168(3): 629–644.
Pummer, T. (2016) 'Whether and Where to Give', *Philosophy & Public Affairs*, 44(1): 77–95.
Pummer, T. (2023) *The Rules of Rescue: Cost, Distance, and Effective Altruism*. Oxford University Press.
——— (2025) 'Compensated Altruism and Moral Autonomy', *Social Philosophy and Policy*, 42(1): 186–203.
Quinn, W. (1989) 'Actions, Intentions, and Consequences: The Doctrine of Double Effect', *Philosophy & Public Affairs*, 18(4): 334–351.
Scanlon, T. M. (1998) *What We Owe to Each Other*. Harvard University Press.
Setiya, K. (2014) 'Love and the Value of a Life', *Philosophical Review*, 123(3): 251–280.
——— (2023) 'Other people', in S. Buss & N. Theunissen (eds), *Rethinking the Value of Humanity*. Oxford University Press.
Shiffrin, S. (1991) 'Moral Autonomy and Agent-Centred Options', *Analysis*, 51(4): 244–254.
Slote, M. (1985) *Common-Sense Morality and Consequentialism*. Routledge.
Slote, M. & Pettit, P. (1984) 'Satisficing Consequentialism', *Proceedings of the Aristotelean Society Supplementary Volumes*, 58: 139–176.
Sung, L. (2022) 'Never Just Save the Few', *Utilitas*, 34(3): 275–288.
Taurek, J. (1977) 'Should the Numbers Count?', *Philosophy & Public Affairs*, 6(4): 293–316.
Tucker, C. (n.d.) 'Autonomy: How Lesser Goods Can Be Justifying Reasons', unpublished manuscript.
Woollard, F. (2013) 'If This is My Body … : A Defence of the Doctrine of Doing and Allowing', *Pacific Philosophical Quarterly*, 94(3): 315–341.
——— (2015) *Doing and Allowing Harm*. Oxford University Press.

3.

Two Kinds of Altruistic Near-Bias

Abstract. May we display near-bias when promoting the welfare of others? I draw a distinction between two different kinds of altruistic near-bias: near-bias within the lives of beneficiaries and near-bias across the lives of beneficiaries. I argue that we generally have moral reasons for exhibiting the latter kind of altruistic near-bias but not the former. This asymmetric result yields some important practical implications. In particular, it suggests that whether or not a 'consistency constraint' governs charitable giving – a constraint that one gives regularly rather than saving at interest to give away greater quantities later – partly depends on the type of charitable institutions to which one is giving.

1. Introduction

It turns out that pigeons are highly impatient.[1] To be precise, pigeons discount the value of future 'Noyes Precision Food Pellets' in a manner predicted by the hyperbolic discounting function:

$$V = \frac{\mathrm{A}}{1 + kD}$$

where V represents the subjective value currently assigned to a future reward A, k is a constant representing degree of impatience, and where D represents length of delay. In

[1] Mazur (1987; 2000).

short: ask a pigeon whether it wants five pellets now or thirty pellets in ten seconds time and it will not care much either way.[2]

In this respect, humans are much like pigeons. We are systematically impatient – or *near-biased* – insofar as we discount the value of more distant future goods. Samuelson's discounting utility (DU) model famously assumes that the agent discounts her future consumption exponentially, such that, across each future period, her concern for some unit of consumption diminishes by the same proportion it diminished over the last.[3] But a large body of evidence in experimental psychology casts doubt on the exponential assumption underpinning DU, instead suggesting that humans generally employ a hyperbolically-shaped rate of discount, just as do pigeons.[4]

Philosophers have perennially debated whether near-bias is prudentially permissible: that is, whether it is rationally permissible to exhibit near-biased attitudes when attempting to promote one's own future good.[5] In the following discussion, I focus on a different question. I ask whether near-bias is *altruistically* permissible: that is, whether it is ethically permissible to exhibit near-biased attitudes with regards to the welfare of others when promoting their good.[6]

My key concern is to show that it is theoretically and practically important to split apart two separate questions hidden within this single question. The questions are:

> *The Intrapersonal Near-Bias Question.* When I am promoting the welfare of a single person, may I be near-biased, such that I intrinsically prioritise (to some degree) her nearer future?

---

[2] Green et. al (2007).
[3] Samuelson (1937).
[4] See Green et. al (1994), Kirby & Herrnstein (1995), and Holt et. al (2008).
[5] In Plato's *Protagoras* (356: b–e), for instance, the suggestion appears to be that prudential near-bias amounts to a failed attempt at the art of 'measurement.' For a survey of more recent debates, see Loewenstein and Elster (1992).
[6] *Inter alia* see Parfit (1984), Broome (1992; 1994), Hare (2008), Purves (2016), de Shalit (2021), Lloyd (2021), Mogensen (2022), Sung (2025).

The Interpersonal Near-Bias Question. When I am promoting the welfare of different persons, may I be near-biased, such that I intrinsically prioritise (to some degree) the welfare of temporally nearer persons?

Drawing together different threads of the literature, I make the case for an as-yet-unstated asymmetry. It is generally impermissible for agents to exhibit altruistic near-bias in the intrapersonal case. But it is generally permissible for agents to exhibit altruistic near-bias in the interpersonal case. Call this the *Asymmetry*$_1$ view.[7]

I argue for the *Asymmetry*$_1$ view in several steps. First, I argue that we generally have substantial moral reasons *against* exhibiting altruistic intrapersonal near-bias. Second, I argue that we generally have substantial moral reasons *in favour* of showing altruistic interpersonal near-bias. The third and final step is to show that the arguments in favour of *interpersonal* near-bias cannot be extended to cover intrapersonal near-bias, and that the arguments against altruistic *intrapersonal* near-bias cannot be extended to cover altruistic interpersonal near-bias. This leaves us with what I have labelled the *Asymmetry*$_1$ view.

Before commencing the argument, it may be worth clarifying the conceptual landscape, and in particular, splitting apart the position I am defending from those positions I am not defending. In broad strokes, the conceptual landscape *vis-à-vis* altruistic near-bias seems to contain at least four distinct possibilities:

Table 1. Possible Models of Altruistic Near-Bias

	Full Altruistic Neutrality	*Full Altruistic Near-Bias*	*Asymmetry*$_1$	*Asymmetry*$_2$
Is altruistic *intrapersonal* near-bias permissible?	✗	✓	✗	✓
Is altruistic *interpersonal* near-bias permissible?	✗	✓	✓	✗

[7] The subscript is necessary – I introduce another asymmetric view momentarily.

For reasons which will quickly become apparent, I find both *Full Altruistic Neutrality* and *Full Altruistic Near-Bias* implausible. That is: I find it implausible that we may *never* exhibit altruistic near-bias, and I find it simultaneously implausible that we may *always* exhibit altruistic near-bias. But one reason why my conclusions may be of particular theoretical interest is that I endorse the *Asymmetry*$_1$ view rather than the *Asymmetry*$_2$ view.

In particular, some theorists might feel that, in endorsing the *Asymmetry*$_1$ view, I have endorsed the 'unnatural asymmetry' rather than the 'natural asymmetry.' For a widespread popular ethical belief appears to rule out the *Asymmetry*$_1$ view while leaving the *Asymmetry*$_2$ view intact: this is the ethical belief that it is generally impermissible to intrinsically prioritise the welfare of certain persons rather than others upon the basis of the timestamp at which they happened to be born.[8] In order to endorse the *Asymmetry*$_1$ view, then, I will have to join with a growing recent movement which attempts to show why this popular belief cannot be used to undermine the ethical legitimacy of interpersonal near-bias in the way that is standardly assumed.

2. Against Intrapersonal Near-Bias

Our very first step is to set out what the phenomenon of intrapersonal altruistic near-bias amounts to. I'll do so with a concrete example. Suppose you can costlessly give some stranger Alice a single utile – an arbitrary improvement to her welfare of fixed magnitude. You must schedule for the utile to arrive in one year's time, in two years' time, in three years' time … and so on. You will be *temporally neutral* with respect to Alice insofar as you are indifferent between these various options. By contrast, you will be *altruistically near-*

[8] See Sidgwick (1907), Rawls (1971), Parfit (1984), and Broome (1992).

biased with respect to Alice insofar as you prefer for the single utile to arrive sooner rather than later. A hallmark feature of such altruistic near-bias is that you will sometimes prefer to give Alice a lesser quantity of utiles in the near-future rather than a greater quantity of utiles in the distant future. The altruistic near-bias in question is *intrapersonal* insofar as it concerns the distribution of benefits within a single life.

We are bracketing-off, here, instrumental grounds for the display of intrapersonal near-bias – grounds such as uncertainty about the future – in order to focus upon what is commonly called the rate of *pure time preference*. All are in agreement, since at least Sidgwick, that further instrumental factors such as uncertainty can clearly justify patterns of near-biased concern.[9] But it is nevertheless important to answer the further question of whether it is permissible to exhibit a *pure* rate of time preference with regards to Alice. After all, the answer to this further question will delimit the maximally permissible extent to which we may in fact show near-bias towards Alice.

As far as starting points go, there is a good argument for thinking that intrapersonal near-bias with regards to Alice is simply ethically impermissible. The argument hinges upon the suggestion that, when all else is equal, an altruist should always prefer the life of the single person she is helping to go better rather than worse:

> *The Maximal Kindness Principle*. When all else is equal – i.e., no other persons are affected and it costs me no more – I should prefer the life of the single person I am helping to go maximally well.

The *Maximal Kindness* principle is highly intuitively plausible: if I am attempting to do benefit a particular beneficiary, and all else is equal, including costs to both myself and to third parties, then I should generally prefer that my beneficiary's life go better rather than worse. It seems, however, that altruists who exhibit intrapersonal near-bias will not

[9] Sidgwick (1907: 381).

generally satisfy the *Maximal Kindness* principle. Such altruists will willingly sacrifice a greater quantity of the good in Alice's later life in order to secure a smaller quantity of the good in Alice's sooner life. The implication of this appears to be that Alice's resulting life ends up being less *overall good.* Insofar as intrapersonal altruistic near-bias violates the *Maximal Kindness* principle, then, we have grounds for rejecting intrapersonal altruistic near-bias.[10]

Of course, many will maintain that the overall goodness of a life does not merely consist in the sum or average of the well-being it contains at different times; in particular, they will maintain that global features about the distribution of welfare *itself* can make a difference to the overall goodness of a life. If this is right then we can't simply assume that giving a beneficiary Alice smaller goods now rather than larger goods later makes Alice's life go overall worse – we can't assume that doing so would violate the *Maximal Kindness* principle. This complication need not undermine the current argument, however. The most famous suggestion about global structure, for instance, is that lives containing a 'rags to riches' structure are better than lives which contain similar amounts of welfare distributed over a 'riches to rags' structure.[11] Agents who exhibit altruistic intrapersonal near-bias with regards to their beneficiaries will not generally promote the putatively valuable rags to riches structure: in being altruistically near-biased, those agents will tend to make the early parts of their beneficiary's life richer and the later parts of their beneficiary's life raggier.[12]

Since intrapersonal altruistic near-bias conflicts with the *Maximal Kindness* principle, which holds independent intuitive plausibility, my own sense is that the burden of proof

[10] This is an other-directed analogue of what Sullivan (2018: 20–33) refers to as the *Life Saving Argument* against prudential near-bias. One natural challenge for the *Maximal Kindness Principle* are considerations of autonomy. I deal with these momentarily.
[11] As in Velleman (1991).
[12] For related complications, see Sullivan (2018: 30–32).

lies on proponents of intrapersonal near-bias to demonstrate why intrapersonal near-bias is in fact morally permissible. Why, then, believe that intrapersonal altruistic near-bias might in fact be ethically permissible? The recent literature contains at least three distinct arguments in its favour. Here, I'll set out these arguments and explain why I find each of them unpersuasive.

(1) The first argument – call it *The Identification Argument* – is developed in very recent work by Leora Sung.[13] The argument proceeds as follows. One first points out that the altruist herself is near-biased with respect to her own interests, and that this may be, in some sense, an inherent part of the lens through which she cannot help but view the world. One then points out that, when altruists bestow benefits upon others, it is usually ethically appropriate for altruists to identify with those they aid such that, in some sense, the altruist feels and appraises those benefits and burdens as her own. It follows from these two claims that it can be ethically appropriate to exhibit altruistic near-bias towards Alice. For I should empathetically identify with Alice as if her own burdens and benefits were my own. And when engaging in such an exercise of empathetic identification, it is ethically permissible *for me* to be especially struck by the importance of near-future goods.

Reply. I find this first argument unpersuasive. The *Identification Argument* would imply, as Sung notes, that I may permissibly exhibit near-bias with respect to my brother's welfare even in the event that he proves to be utterly temporally neutral. Overall, though, this implication is hard to believe. In particular, it seems that this would be to disregard my brother's preferences and substitute them for my own in a manner broadly reminiscent of paternalism. Indeed, such a substitution may even fail to qualify as paternalistic in the usual sense, since there is no guarantee that such a substitution will be conducive to my brother's overall good.

[13] Sung (2025). My thanks to Leora Sung for clarifying some of the relevant suggestions in what follows.

Sung remarks that the such a suggestion may not be unintuitive, and that it is a natural reflection of the idea that "we should treat other people the way we would want to be treated."[14] But this thought risks eliding a subtle indexical distinction. It's not plausibly true, as a general rule, that if I desire x I should thus care that *you* get x. It's more plausibly true, as a general rule, that if I desire x and you desire y, I should care that you get what *you* desire (y) just as you should care that I get what *I* desire (x). Spelled out in these terms, it is the former and more controversial reading needed in order to vindicate the suggestion developed here by Sung.[15]

(2) The second argument – call it *The Autonomy Argument* – makes reference to the notion of autonomy.[16] In brief, if Alice is an ordinary human being then she will be near-biased with respect to her own well-being. But if Alice is near-biased with respect to her own well-being, then you ought to be near-biased in your attempts to benefit Alice. After all, this is what a concern for Alice's autonomy demands of you. A concern for Alice's autonomy demands that you benefit her in the manner in which she would prefer to be benefitted, and not benefit her according to your own strictures.

Reply. The *Autonomy Argument* is more promising, but it nevertheless goes awry. In particular, the *Autonomy Argument* presupposes that an adequate concern for the autonomy of some beneficiary x involves deference to the temporal attitudes which x currently holds. This is not obvious, however. If, for instance, Alice holds temporal attitude a_1 at time t_1 (an overriding concern for t_1 goods) and I know that she will hold a conflicting temporal attitude a_2 at t_2 (an overriding concern for t_2 goods), then it is not obvious why a concern

[14] Sung (2025: 379).
[15] George Bernard Shaw puts the same concern in quippier terms: "Do not do unto others as you would that they should do unto you. Their tastes may not be the same."
[16] This argument has origins in Hare (2008), who suggests that appropriate concern for Alice requires imaginatively engaging with her circumstances in order to take onboard those temporal preferences which Alice actually holds.

for *Alice's* autonomy dictates deferral at t_1 to her t_1 attitude rather than to her t_2 attitude. Real-world altruists do indeed find themselves in a case with such a structure: since beneficiaries are typically near-biased at each point in time, an altruist can know that adopting near-biased preferences with respect to Alice's welfare expressly conflicts with those near-biased preferences Alice will very shortly come to hold.[17]

Of course, it is easy to claim that, in such contexts, a concern for Alice's autonomy simply demands that I defer to Alice's present temporal attitudes rather than to Alice's future temporal attitudes, for Alice's present attitudes are those attitudes which she hold contemporaneous to my decision. But this move presupposes precisely that temporal attitudes located in Alice's present have priority over temporal attitudes located in Alice's future, which is the very claim we hoped that the autonomy argument could demonstrate for us in the first place. It's thus unclear whether the appeal to autonomy can support the case for altruistic near-bias in a manner which is non-circular.[18]

One further point bears mentioning. If Alice is a normal human, then she will tend to discount her future welfare hyperbolically.[19] It is a well-discussed result that agents who discount their future welfare hyperbolically will be liable to exploitability – that is, they will willingly accept a series of trades which renders them worse-off at each point in time and better at none.[20] An altruist who adopts Alice's temporal attitudes in dealing with

---

[17] Of course, there would not be a tension between present and future Alice if Alice employed an exponential (constant) discount rate, but as the evidence in fn.4 suggests, people do not in practice employ an exponential (constant) discount rate.

[18] As Bykvist (2003) notes, it often seems appropriate to screen-off *past* preferences which one's beneficiary has now outgrown or discarded. It's unclear the extent to which this insight, if correct, would bear upon my current remarks. If the view is simply that past preferences ought to be screened off because the satisfaction of past preferences does not bear on autonomy, we would face the same challenge: why should the satisfaction of past preferences, if they were sincere and properly formed, not be understood to bear on autonomy? If the view is that past preferences ought to be screened off because those preferences are not the beneficiary's 'final word' (i.e., they are preferences which the beneficiary will outgrow in the future), then the view would not resolve the present challenge. The view would not support conformity with Alice's (present) t_1 temporal attitudes rather than her (future) t_2 temporal attitudes. For just as Alice's present temporal attitudes outgrow her past ones, her future temporal attitudes will outgrow her present ones.

[19] See again fn.4.

[20] The *locus classicus* being Strotz (1955).

Alice will thus be liable to exhibit *altruistic exploitability*: she will be liable to do what is worse for Alice at each point in time and what is better for Alice at none.[21]

Importantly, if Alice knew that her persistent near-bias rendered her vulnerable to exploitability of such a kind, she might come to feel rather differently about her temporal near-bias. She might, for instance, come to deny that her near-bias can be identified with her 'true self.'[22] She might desire that she could drop those near-biased attitudes, even if she felt that, in some sense, she lacked the psychological ability to do so. And most importantly, she might deny that her own temporal attitudes should serve as a guide to how others ought to treat her. If Alice's near-biased preferences held these qualities – if they were 'shallow' in these various salient ways – then it becomes far less obvious that a concern for Alice's autonomy mandates that altruists adopt Alice's current (near-biased) temporal attitudes as a guide to her treatment.

(3) The third argument – call it the *Empathy Argument* – has been developed at greatest length in the work of Michael Slote.[23] The key starting idea here, already familiar to ethicists of care, is that the concept of empathy is somehow of fundamental importance in attempting to understand and interpret the moral realm – where empathy involves, in very rough terms, agents having the capacity to affectively feel the goods and bads of others, rather than merely knowing, in a descriptive manner, that such goods and bads exist.[24]

---

[21] A concrete example may help, adopted from the prudential example given by Ahmed (2020). Suppose that, in any given calendar year Y_i, Alice is indifferent between *0.2* lollipops in Y_{i+1} versus *1* lollipop in Y_{i+2}. But in any given year Y, she is indifferent between *0.8* lollipops in Y_{i+2} versus *1* lollipop in Y_{i+3}. Suppose you adopt these non-constant temporal attitudes when benefitting Alice. Then you will be liable to *altruistic exploitability* on Alice's behalf. In the first year you will willingly shift Alice from a four-year lollipop schedule of (0, 0, 60, 60) to (0, 0, 40, 90), since you care about gaining 30 lollipops in three years' time more than you care about losing 20 lollipops in two years' time. But in the second year you will willingly shift Alice from the lollipop schedule (0, 0, 40, 90) to (0, 0, 50, 50), since you care about securing Alice 10 lollipops in one year's time more than you care about losing 40 lollipops in two years' time. You will thus benefit Alice *less* at each point in time in which she is benefitted.

[22] As in Frankfurt (1971).

[23] Slote (2003; 2007).

[24] Slote (2003: 180).

As Slote observes, it might be possible for the empathy theorist to justify altruistic near-bias in the following way. We usually feel greater empathy for those in our immediate spatial and temporal vicinity. But it is one of the central insights of empathy theorists that feeling empathy towards a patient – or perhaps, being situated such that I *can* feel empathy towards that patient – justifies showing that patient a special kind of moral concern. This gives us all the tools we need for drawing the relevant conclusion: it is justifiable to show a special moral concern for Alice's immediate rather than more distant future, given that I can feel greater empathy towards near-future Alice than I can towards distant-future Alice.[25] In this way, considerations of empathy might be capable of grounding near-biased preferences with respect to Alice's flourishing over time.

Reply. I am happy to join with care and empathy theorists in emphasising the moral importance of empathy relations. But a key challenge for the *Empathy Argument* lies in the recent experimental data.

Now, the recent experimental data does indeed provide strong support for the phenomenon of *intertemporal empathy decline*, defined by Coleman and DeSteno as "the tendency to feel less empathy for others' suffering in the future than equivalent suffering in the present."[26] For instance, when told of a harm event *h* befalling a present stranger Mary versus Mary in 25 years' time, participants tend to exhibit greater levels of empathetic distress response with regards to present (rather than future) Mary.

But unfortunately for the *Empathy Argument*, a crucial lesson borne out by the recent experimental data is that the above result doesn't indicate any kind of natural incapacity of agents to feel equal levels of empathetic distress for present versus future Mary. Even extremely brief 'vividness interventions' in which participants momentarily reflect upon

[25] *Ibid.*
[26] Coleman and DeSteno (2024: 1976).

the plight of future Mary result in participants showing equalised levels of empathetic distress with regards to present versus future Mary.[27]

This result seems to suggest that any empathy-based argument for intrapersonal near-bias is on shaky grounds. It would be curious to claim that my empathetic reactions towards Mary justify a special treatment of near-future Mary, if, given only (literal) seconds of internal reflection, I will come to feel the *same* level of empathetic distress towards future Mary as I do towards present Mary.

One way of putting the objection is that the *Empathy Argument* neglects the fact that, at least with regards to a contemporary like Mary, we tend to have a certain kind of mental control over whether or not we feel the same kind of empathy towards Mary's future suffering as we do towards her present suffering. We can very easily exercise this mental control to fully empathise with Mary's future suffering as we do with her present suffering. If the possibility of equal empathetic distress is within such easy reach, it seems far too strong to say that one's initial empathetic reaction towards present Mary mandates (or even makes permissible) uniquely special moral concern for present Mary.

3. In Favour of Interpersonal Near-Bias

I've suggested that, in general, we ought not display intrapersonal altruistic near-bias. What, though, about *interpersonal* altruistic near-bias? As before, the first step here is to spell out what the phenomenon of interpersonal altruistic near-bias amounts to.

Suppose that you are in generation g_1. You can give a single utile to Beatrice in generation g_2, to Charlie in generation g_3, or to Daniel in generation g_4. (Idealise that none of the generations overlap.) You will be temporally neutral with respect to future persons

[27] Coleman and DeSteno (2024: 1981–1982).

insofar as you are indifferent between these various options. By contrast, you will be altruistically near-biased with respect to future persons insofar as you prefer to give the single utile to Beatrice. A hallmark feature of such altruistic near-bias is that you will sometimes prefer to give Beatrice a lesser quantity of utiles rather than giving (say) Daniel a greater quantity of utiles. The near-bias in question is *interpersonal* insofar as you are near-biased in your distribution of benefits across separate and temporally subsequent persons. As before, we bracket-off instrumental factors such as uncertainty in order to focus upon the permissible rate of pure time preference.

The very first thing to say here is that, already, the case for *interpersonal* altruistic near-bias is on firmer ground than the case for *intrapersonal* altruistic near-bias. This is because displaying interpersonal altruistic near-bias seems compatible with the *Maximal Kindness* principle: the principle that, for any given beneficiary, holding all other factors equal, including our treatment of other persons, we should prefer to make that beneficiary's life go better rather than worse. Interpersonal altruistic near-bias is compatible with the *Maximal Kindness* principle because interpersonal altruistic near-bias involves a decision maker making trade-offs as they benefit different temporally subsequent persons.

Of course, displaying interpersonal altruistic near-bias might conflict with other general principles. Displaying interpersonal altruistic near-bias might violate a general principle that one should always aim to maximise the impersonal sum of welfare across persons, for instance. But I take it that this is just the sort of further substantive principle towards which many non-consequentialists, for instance, won't feel pulled in the first place. At the very least, the challenge from *Maximal Kindness* seems to affect intrapersonal near-bias 'out of the gate' in a way that it does not so clearly affect interpersonal near-bias.

Now, I previously suggested that many of the arguments in favour of *intrapersonal* altruistic near-bias were unsuccessful. By contrast, it strikes me that the arguments in favour of *interpersonal* altruistic near-bias tend to hold greater plausibility.

The most promising argument, here, is what we might call the *Partiality Argument*. The key idea behind the partiality argument is that we have agent-relative partial reasons, of some kind or another *P*, which mandate that we show special moral concern to the those persons who will find themselves born in the nearer (rather than more distant) future. Such reasons are reasons that we hold towards nearer-future persons precisely in virtue of the distinctive position in which they stand relative to us, and in which we stand relative to them.

What might the *P*-reasons be? The most detailed account of candidate *P*-reasons is given in recent work by Andreas Mogensen.[28] Mogensen draws attention to the fact that nearer-future generations typically contain persons who are more closely our kin, and that we typically have agent-relative partial reasons for showing special moral concern towards those who are more closely our kin. After all, in general, it seems that our agent-relative partial reasons for favouring our children are stronger than our agent-relative partial reasons for favouring our grandchildren, which are stronger than our agent-relative partial reasons for favouring our great-grandchildren … and so on. In this way, the *P*-reasons might turn out to support a near-biased pattern of altruistic concern across generations.

The kinship argument gains more traction in some decision contexts than in others. Suppose, for instance, that a policy maker in some country *C* can benefit future *C*-persons on behalf of current *C*-persons, and that the typical present *C*-person expects to be the ancestor of future *C*-persons. Such a policy maker might take herself to have partial

[28] Mogensen (2022).

reasons (of kinship) to benefit more proximate *C*-persons, where these reasons would consist in the collection of individual kinship reasons held by those present persons on whose behalf she acts.[29] By contrast, there are other contexts in which the kinship argument gains less traction. Suppose that I *qua* individual have a chance to donate to a random person within generation g_3 versus g_5, and that I do not expect to have any descendants at all – even obliquely via my siblings or cousins, say. In that case, it is much harder to see how the kinship argument might generate a case for interpersonal near-bias.

It's worth emphasising, however, that kinship reasons are not the only candidate *P*-reasons. Thus, if there are contexts in which kinship reasons are absent, this need not imply that *no other P*-reasons could justify interpersonal near-bias. What Harry R. Lloyd recently observes, for instance, is that we more closely share a *scope of moral concern* with near-future persons than we do with more distant future persons.[30] There is, for instance, a greater overlap in the scope of 'moral things' that you and I care about and are capable of influencing than there is between myself and a stranger who will come to be born five-thousand years from now. Lloyd points out that, to the extent that I share a distinctive scope of moral concern with nearer-future generations, this may be just the sort of consideration which justifies greater partial concern for the welfare of nearer-future generations.[31]

For what it's worth, I think that there may be even more potential *P*-reasons than kinship reasons and shared scope reasons, where these further potential *P*-reasons may,

[29] See discussion in Mogensen (2022: 2742–2743). Prior to Mogensen's discussion of kinship, economists were aware of the relevance of partiality to these debates, as in Stern (2008).
[30] Lloyd (2021).
[31] An immediate reaction is that Lloyd's argument could be used to justify substantive *intrapersonal* altruistic near-bias, since I share a greater moral scope with 'present' Alice than I do with 'future' Alice. I return to this concern in §4.2; my thanks to Harry R. Lloyd for emphasising these possibilities.

in a similar way, justify near-biased patterns of intergenerational concern. The following possibility strikes me as particularly promising.

Recall that I previously rejected the *Empathy Argument* in favour of intrapersonal altruistic near-bias. I rejected this argument on the grounds that it is comparatively easy for us to feel the same level of empathetic distress that we feel towards present slices of Mary as we do towards future slices of Mary. This is plausibly true when we focus our attention on any given contemporary who lives a life much like ours. But it strikes me that things may change, in important and evaluatively relevant respects, when we shift our attention to focus upon generations spread out across the very distant future. One salient possibility, for instance, is that beneficiaries in very distant future generations will live lives which are *alien* or *conceptually incomprehensible* when compared with our own. We may be capable of making general judgements that such distant-future lives could be good or bad, but simultaneously find ourselves unable to grasp the qualitatively diverse ways in which those distant-future lives could be good or bad. Perhaps the Romans were in the same position with regards to us: the Romans, *sans* any Doctor-Who-style time travellers, did not live out their lives with access to the cognitive categories which would have allowed them to make sense of the precise ways in which the internet and international air travel have been good and bad for human flourishing.

What we might say, here, is that the plight of more proximate generations is generally *relatable* to us in a way that the plight of distant generations is not, where relatability is a matter of having the cognitive and conceptual categories needed in order to make sense of the precise ways in which certain experiences can be good and bad for us. We could then say that the relation of relatability serves as potential grounds for showing partial concern towards this-or-that beneficiary: the fact that I can relate to the suffering of

another person, in this distinctive way, serves as a grounds for showing partiality towards their plight.[32]

Those, then, are three credible ways in which we can derive *P*-reasons which would justify the display of special concern towards more proximate generations. I'll say more about these three strategies shortly. For now, it's worth emphasising just a few important points. First, as I've already suggested, these strategies aren't mutually exclusive: it's perfectly comprehensible, for instance, to maintain that both kinship reasons and shared scope reasons have some distinctive role to play in justifying interpersonal near-bias. Second, it's nonetheless plausible that there are firm limits on the extent to which these reasons can justify interpersonal near-bias: even if *P*-reasons can justify some level of special concern for more proximate generations, it's plausible that we always owe some substantive and non-trivial level of moral concern towards future generations, regardless of their futurity. The third point to emphasise is that, at least in part, these arguments turn out to favour interpersonal (rather than intrapersonal) altruistic near-bias as a result of the contingent facts governing human biology and social life. If, for instance, humans lived for 10,000 years, then perhaps I would have *P*-reasons of relatability to favour near-future Mary rather than distant-future Mary, where these reasons would be grounded in the fact that distant-future Mary lives a life which is conceptually incomprehensible to my present self. This shouldn't be seen as a weakness in the argument: my aim isn't to demonstrate that the *Asymmtery*$_1$ view is necessarily true, but only, that it has plausibility contingent on the facts which happen to constrain human life.

[32] I've since learned that Riedener (2025) makes related suggestions in a very recent contribution. Now, one might worry: couldn't rapid advances in technology cause human life to become conceptually incomprehensible over a period of mere decades? And couldn't this ground *P*-concern, in the intrapersonal case, for present rather than future Alice? I take the point, since I don't mean to carve out a dogmatic distinction according to which intrapersonal altruistic near-bias could *never* be justified. Nevertheless, I remain reasonably sceptical: we clearly have a much better grasp of the shape which human life might take in several decades than of that it might take in (say) several millennia.

4. Objections

I've so far built the following case: intrapersonal altruistic near-bias is generally impermissible, but interpersonal altruistic near-bias is generally permissible. I still have important work to do, though. The next step of the discussion is to show that some of the objections against intrapersonal near-bias can't be extended (with much plausibility) against interpersonal near-bias; similarly, I'll argue that some of the arguments for interpersonal near-bias can't be extended (with much plausibility) to cover intrapersonal near-bias. Without this final step the argument for the *Asymmetry*$_1$ view risks collapsing to simply an argument for *Full Altruistic Neutrality* or *Full Altruistic Near-Bias*.

4.1 An Objection to Interpersonal Near-Bias

First, let's turn back the clock to my discussion of intrapersonal near-bias. In my discussion of intrapersonal near-bias I briefly noted the following point. If you discount Alice's future welfare in a non-constant (hyperbolic) manner, then you will be liable to exhibit *altruistic exploitability* with regards to Alice's welfare. That is, you will willingly accept trade-offs such that you do what is worse for Alice sometimes and better for her never.[33]

But if we take the above partiality arguments in favour of interpersonal near-bias seriously, it seems that we will be obliged to discount the welfare of future generations in a non-constant (hyperbolic) manner. One crucial reason why our discounting must be non-constant (hyperbolic) is that a constant rate of discount would imply that our concern for future persons peters-out to zero in the long-run: as already mentioned, this is hard to

[33] As previously, see Strotz (1955).

believe, and conflicts with the appealing idea that we always owe future persons some substantive minimal level of concern regardless of their futurity.[34]

This seems to suggest that, if we engage in interpersonal near-bias, we will be vulnerable to *intergenerational altruistic exploitability*: that is, we will willingly accept a series of trade-offs such that we do what is worse for some generations and better for none. One might reasonably worry that this result threatens to undermine the ethical legitimacy of interpersonal near-bias altogether.

A concrete example will help illustrate the point. Consider:

> **Lollipops.** Starting a couple of generations from now, every future generation will receive sixty lollipops. Each generation contains a 'deciding altruist' who can modify the schedule of future lollipops. Each 'deciding altruist' is near-biased. For any given generation G_i, the deciding altruist of that generation is indifferent between giving a G_{i+1} member *0.2* lollipops and a G_{i+2} member *1* lollipop. But the deciding altruist cares far less about delays which take place in the distant future. For any given generation G_i, the deciding altruist of that generation is indifferent between giving *0.8* lollipops to a G_{i+2} member versus *1* lollipop to a G_{i+3} member.[35]

Suppose, in **Lollipops,** that you are the first generation's deciding altruist. This means you initially anticipate the following schedule of intergenerational benefits:

G_1 receives 0.
G_2 receives 0.
G_3 receives 60.
G_4 receives 60.

Given your non-constant rate of discount, you will willingly forfeit 20 G_3 lollipops in order to secure 30 G_4 lollipops. That is, you will willingly trade to:

[34] As originally noted in Parfit (1984: 485–486).
[35] As in fn.21, this example is loosely adopted from the intrapersonal example in Ahmed (2020: 248).

G_1 receives 0.
G_2 receives 0.
G_3 receives 40.
G_4 receives 90.

When you are dead and buried, however, generation G_2's deciding altruist will willingly forego 40 G_4 lollipops in order to secure 10 G_3 lollipops. That is, they will willingly trade to:

G_1 receives 0.
G_2 receives 0.
G_3 receives 50.
G_4 receives 50.

As a result, each benefitted generation is left worse-off and no benefitted generation is left better-off. In light of this fact, the objection goes, it's hard to believe that interpersonal altruistic near-bias (underpinned by the aforementioned partiality arguments) could have ethical legitimacy.

I don't have anything especially novel to say about the above objection from exploitability. It seems to me, broadly speaking, that the objection from exploitability lacks force when we recall the intergenerational context in which it is currently being made. What distinguishes the intergenerational context is the presence of multiple independent altruists whose different perspectives cannot, to use a phrase of Mogensen's, be easily grouped as the contents of a "suitably unified outlook."[36] The objection from exploitability only strikes me as having substantial force when the multiple perspectives at work can indeed be grouped as the contents of a suitably unified outlook. In saying that the relevant perspectives here cannot be grouped as the contents of a suitably unified outlook, I mean roughly the following. In the intergenerational context, exploitation

[36] Mogensen (2022: 2739).

comes about as a result of separate persons governed by separate sets of agent-relative reasons responding to separate sets of moral priorities. In this sort of disunified context, it is unclear to me whether the objection from exploitability should be viewed as having much force.

By way of illustration, consider:

> **Family Feud**. You and I are both parents whose respective children have been up to no good. Now the police are knocking on our doors and want answers. If we both keep silent, each child gets five years. If we both confess, each child gets ten years. But if I confess and you stay silent, my child goes free and yours gets fifteen years. *Mutatis mutandis* for you. The case is non-repeatable, lacks the possibility of communication … and so on.[37]

Family Feud is a case in which (I, you) face the following set of agent-relative incentives:

Table 2. Agent-Relative Incentives in Family Feud

	You keep silent	You defect
I keep silent	(-5,-5)	(-15,0)
I defect	(0,-15)	(-10,-10)

If each of us displays an agent-relative concern for the welfare of our children then we will each defect. This is because (1) from my distinctive perspective defecting dominates whatever else you do, and because (2) from your distinctive perspective defecting dominates whatever else I do. The natural upshot is that we together choose a socially suboptimal outcome. But insofar as one takes seriously a category of agent-relative reasons in the first place, it's hard to see why this should serve as a *reductio* of such reasons, or proof that it's therefore morally wrong to act partially upon those reasons. An alternative interpretation is simply that the problem lies in the world rather than within us: you and I face a tragic set of circumstances in which our agent-relative practical

[37] Such cases are familiar from Parfit (1984: 96–108). For relevant recent discussion, see Fullhart (2023).

reasons are fundamentally misaligned, and in which neither of us can hope to pursue what is best, given how we expect the other to act.[38]

Insofar as this is our sentiment concerning **Family Feud**, then an analogous point can be made in the intergenerational case. Insofar as we take seriously a category of partial and agent-relative reasons in the first place, the intergenerational case is one in which the partial and agent-relative reasons held by competing agents across generations simply turn out to be fundamentally (but tragically) misaligned. This opens up the possibility that, sometimes, we are unable to pursue what would be best, conditional on how we expect other generations to act.

4.2 An Argument for Intrapersonal Near-Bias

I vindicated intergenerational near-bias on the grounds that certain kinds of *P*-facts – kinship facts, shared moral scope facts, and relatability facts – seem to ground the display of agent-relative partial concern for the welfare of those in more proximate generations. A very natural thought, however, goes as follows. Couldn't we extend these partiality arguments such that they vindicate the permissibility of *intrapersonal* altruistic near-bias too?

We would need to do some novel thinking, here, to figure out how the reasoning might go. Here is one way in which we could develop the thought. Perhaps I do indeed have stronger partial and agent-relative reasons of kinship for favouring (say) my near-future brother rather than my more distant-future brother. Perhaps this result could be grounded in the fact that my near-future brother is, in some sense, *more my brother* than is my distant-future brother. Similarly, perhaps I do indeed have stronger partial and agent-

[38] Of course, Parfit's (1984: 96–108) own interpretation is that, with qualifications, such cases play a role in defeating the commonsense view that morality contains partial dimensions.

relative reasons of kinship for favouring (say) an as-yet unborn descendant at the start of her life rather than at the end of her life. Perhaps this could be grounded in the fact that the descendant in question would be *more my great-niece* at the start of her life than she would be at the end of her life.

The very first thing to say about this strategy is that it only seems to work for some of the partiality arguments I set out previously. As may already be apparent, it seems a large stretch to try and make this argument work with respect to considerations of kinship. The commonsense view, I take it, is that my future brother is just as much my brother to present me as is my present brother, and that, insofar as I have partial reasons for treating my brother especially well, then these apply equally well to 'future brother' as they do to 'present brother.' To deny as much would be to ask us to rethink much of our ordinary moral thought and talk.

Perhaps other species of the partiality argument would fare better. Consider again the partiality argument developed by Lloyd. One could perhaps argue that just as a 'shared scope of moral concern' justifies a special partial concern for more proximate generations, it justifies a special partial concern for the more proximate time-slices of (say) my neighbour Alice. Future Alice, for instance, has a lesser shared scope of moral concern with my present self than does present Alice: there are fewer 'moral things' which both future Alice and myself share the ability to influence.

This is a more promising avenue. But the key point to make here is that, even if right, in practice, this objection may only ground a negligible or extremely limited form of intrapersonal discounting – not at all the sort of discounting which would have previously been predicted by, say, the *Autonomy Argument* or the *Identification Argument.*[39] The key

[39] For instance, both the *Autonomy Argument* and the *Identification Argument* would ground extremely rapid discounting of patients' futures. Hausman (1979), to give one famous example, observes that individual consumers buy cheaper (but shoddier) air conditioners in the manner predicted by an 89% annual

reason is that this objection grounds a permissible intrapersonal discount rate for our contemporaries in the same core facts of partiality – decaying 'shared scope of moral concern' – which grounded the interpersonal discount rate.

Thus, consider the case of an individual beneficiary p_i born at roughly the moment of my choice and scheduled to die many happy years later:

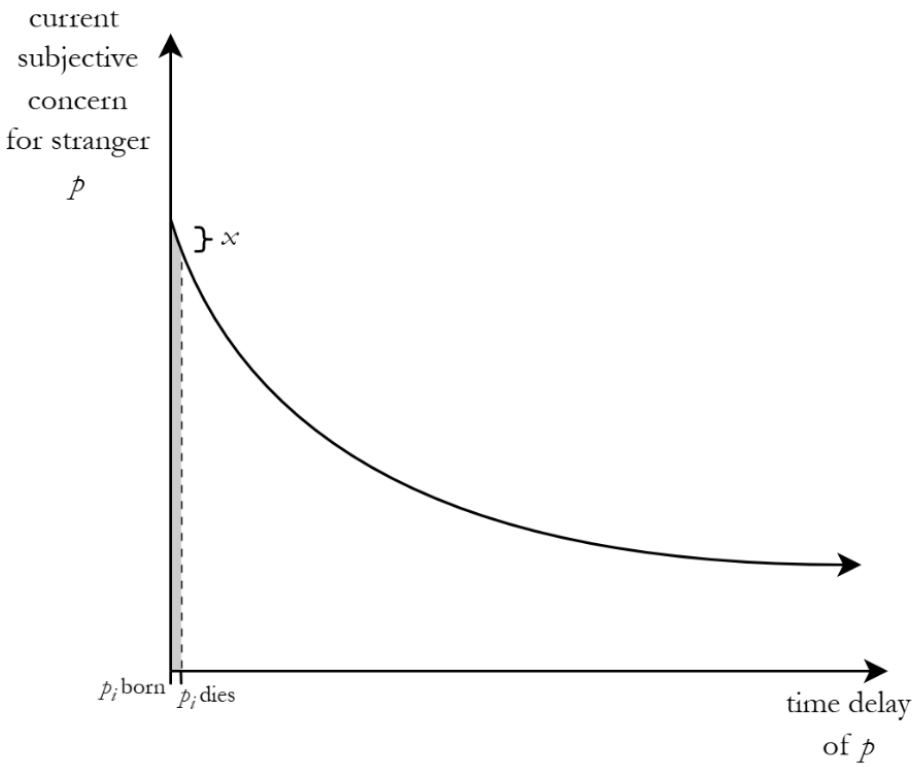


Fig.1 Subjective concern for future persons, including p_i

The entire life of p_i is captured by the shaded zone. The important point to make here is that, insofar as I may currently discount p_i's future time slices on partial and agent-relative grounds, and insofar as these partial and agent-relative grounds are the same grounds which permit discounting across generations in general, then the extent to which I may discount p_i's future relative to her present will be only extremely limited – as given by the

prudential discount rate. This, then, is the sort of discount rate which the *Autonomy Argument* and the *Identification Argument* would seem to require.

measure x. Indeed, a high level of intrapersonal discounting for p_i will require a rapid decrease in subjective concern for future persons in general. And this seems to imply that we should discount p_i's future in an extremely limited – perhaps practically negligible – manner. For presumably, through all of this, we wish to preserve the verdict that persons born in the near-future are owed a substantive and robust level of moral concern.

5. The Consistency Constraint

That concludes my defence of the *Asymmetry*$_1$ view. I've suggested that we generally lack moral reasons for engaging in intrapersonal near-bias. But we generally possess moral reasons for engaging in interpersonal near-bias. In closing, I'll show how this result sheds light on certain recent debates concerning the nature of our obligations of beneficence.

A suggestion frequently made within the non-consequentialist literature on beneficence is that giving should be governed by a *consistency constraint*. The suggestion, here, is something along the following lines: there is a moral imperative to give regularly (i.e. monthly or yearly), sacrificing our financial gains whenever we come into them, rather than hoarding up some sufficiently large quantity in order to give it away upon our deathbeds. This is even if hoarding wealth in a long-term investment account with competitive interest would allow us to do *more good later*.[40]

Suggestions broadly along these lines arise in various contexts. Related suggestions are often made by those who view beneficence as an imperfect duty: amongst those who believe that, broadly speaking, beneficence is an end of our lives but not an end which there is a determinate requirement to pursue at all times. Thus, Thomas Hill, Jr. writes that we must set the happiness of others not just as an end, but as a "continually relevant"

[40] For extended recent discussion, see Sung (2025).

end.[41] And Robert Noggle commits to the view that the moral agent must have or cultivate a propensity to promote the end of human flourishing, where it is evidence of a failure of a propensity to promote an end if there is an extended series of opportunities in which I fail to promote or attend to that end.[42]

Garrett Cullity's early remarks concerning the limits of beneficence also gestured in this direction, in suggesting that agents ought to calibrate the limits of beneficence by considering the amount they gave within the previous year: the practical upshot of this calibration strategy, it seems, is that agents are 'penalised' in the event that they have not engaged in regular recent giving (i.e., insofar as the agent gave a thousand dollars a year ago rather than a thousand dollars *across* the last year).[43] And recently, in his specification of the requiring versus permitting reasons framework, Theron Pummer raises the possibility that the strength of one's permitting reasons against sacrifice at some time t could be weaker insofar as one's past altruistic sacrifices of magnitude n came further in the past, or insofar as one's future sacrifices of magnitude n will come further in the future.[44]

Of course, these various remarks are diverse. For now, let's simply idealise that the consistency constraint has the following general form: agents should give regularly rather than hoarding to give later, even if this results in agents giving less money (and thus producing less good) overall, by some suitable margin.

The important point to make here is that at least one possible way of explaining and justifying the consistency constraint would involve citing considerations of altruistic near-bias. To see as much, consider the following case. Imagine you can greatly benefit

[41] Hill, Jr. (2002: 206).
[42] Noggle (2009: 7).
[43] Cullity (2003: 416).
[44] Pummer (2023: 143).

some stranger Bernard: every time you sacrifice one unit of your own welfare, you simultaneously improve Bernard's welfare by ten units. You then discover that you may benefit Bernard according to one of the following two schedules:

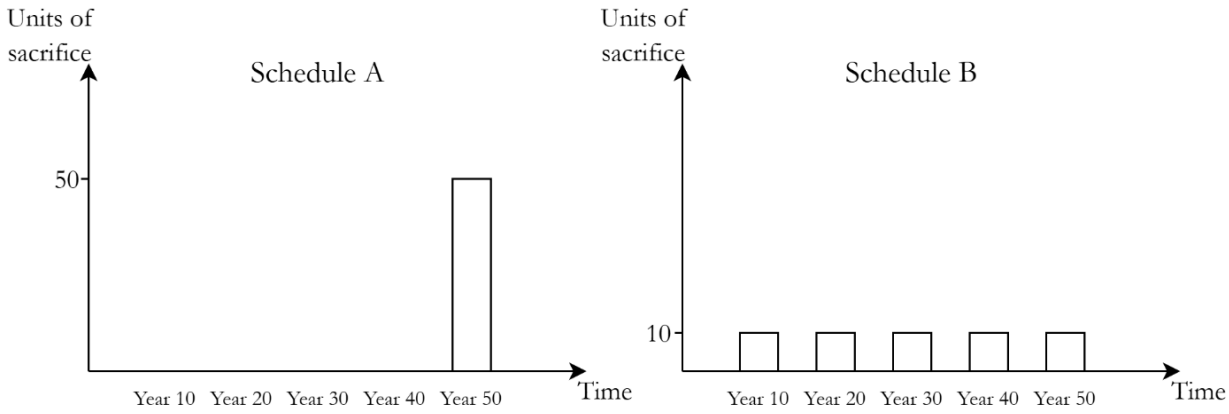


Fig 2. Two Sacrifice Schedules

If it is appropriate to be near-biased with respect to Bernard's welfare, then it turns out that you have moral reason for preferring the 'consistent' schedule of sacrifice – that is, for preferring Schedule B. Why is this? Well, simply put, insofar as Schedule B involves consistently benefitting Bernard, it involves bringing certain benefits to Bernard *sooner rather than later*. And this is something of which the near-biased altruist approves. The near-biased altruist thinks that we *should* do our best to benefit Bernard sooner rather than later.[45]

[45] Of course, if you are near-biased, then the ideal possibility seems to involve giving away an enormous lump sum to Bernard right now – a possibility picked out by neither Schedule A nor Schedule B. The point is duly taken. But note that, in an extremely wide range of practically relevant cases, givers are indeed faced with an exclusive choice between Schedule A and Schedule B. These are those altruistic contexts in which decision makers can either give away some portion of their (say) regular monthly pay cheque or instead invest their pay cheque with interest to donate greater quantities later. Altruistic near-bias would suggest that, in these commonly discussed choice contexts, there is indeed reason to give away a portion of one's income the moment it is received, even given the possibility saving with long-term interest. (For recent discussion, see Sung (2025).)

If this is right, then my previous conclusions about altruistic near-bias are going to have implications for the legitimacy of the consistency constraint. I've so far argued that we should not engage in intrapersonal altruistic near-bias: I've argued, in particular, that it is very difficult to justify the claim that we should be near-biased *vis-à-vis* the welfare of any given beneficiary like Bernard. This means, in turn, that it will be much harder to ground a consistency requirement on altruistic giving *vis-à-vis* any given beneficiary like Bernard. All else equal, we should exhibit something much closer to indifference when confronted with Schedule *A* versus Schedule *B*.

But of course, as I have been advocating for the *Asymmtery*$_1$, there may nonetheless be other contexts in which the consistency requirement will gain more traction. For instance, suppose that, in choosing between Schedule *A* and Schedule *B*, my altruistic donations don't benefit Bernard. They will instead benefit a random person within the generation to be born thirty years after my act of donating – perhaps, as is not impossible, by helping to fund medical infrastructure which will take several decades to build.

In such a case, Schedule *A* concentrates benefits upon future generations – more precisely, upon those persons who will find themselves born thirty years after my death. By contrast, Schedule *B* distributes benefits across both present generations *and* future generations. Insofar as we have special agent-relative reasons to exhibit near-bias with respect to the welfare of those born in more proximate generations (kinship reasons, for instance) then we will have special grounds for preferring Schedule *B* over Schedule *A*. This could be the case even if Schedule *B* secured less overall good by some suitable margin *n*.

For what it's worth, I find these results regarding the consistency constraint reasonably intuitively plausible. In the interpersonal case, I feel that I have special moral reasons to start my giving now. In the intrapersonal case, I feel it may matter less whether

I give to Bernard now versus later. Insofar as these results have intuitive appeal, they may serve to vindicate the central asymmetry I've drawn in these pages.

6. Conclusion

Should we be near-biased altruists? My suggested answer to this question is: in one way but not another. We ought not engage in altruistic intrapersonal near-bias. But we ought to engage, to some degree, in altruistic interpersonal near-bias.

The above discussion has been one step forward in helping us to figure out the exact relationship which might hold between intrapersonal versus interpersonal altruistic near-bias. Even for those who disagree with my preferred conclusion, the above discussion has nonetheless served the following useful purpose: the above discussion brings to light some of the distinctive ethical considerations at play within either altruistic context, showing, in particular, that the distinctive ethical considerations which justify one kind of near-bias need not, as a conceptual matter, justify the other.

The discussion has also yielded some important implications for our obligations of beneficence. Is charitable giving governed by a consistency constraint? My suggested answer to this question too: in one way but not another.

References

Ahmed, A. (2020) 'Rationality and Future Discounting', *Topoi*, 39(2): 245–256.

Broome, J. (1992) *Counting the Cost of Global Warming*. White Horse Press.

——— (1994) 'Discounting the Future', *Philosophy & Public Affairs*, 23(2): 128–156.

Bykvist, K. (2003) 'The Moral Relevance of Past Preferences', in H. Dyke (ed), *Time and Ethics: Essays at the Intersection*. Springer.

Coleman, M. and DeSteno, D. (2024) 'Intertemporal Empathy Decline: Feeling Less Distress for Future Others' Suffering', *Emotion*, 24(8): 1975–1984.

Cullity, G. (2003) 'Asking Too Much', *The Monist*, 86(3): 402–418.

Frankfurt, H. (1971) 'Freedom of the Will and the Concept of a Person', *The Journal of Philosophy*, 68(1): 5–20.

Fullhart, S. (2023) 'Embracing Self-Defeat in Normative Theory', *Philosophy and Phenomenological Research*, 109(1): 204–225.

Green, L., Fristoe, N., & Myerson, J. (1994) 'Temporal Discounting and Preference Reversals in Choice Between Delayed Outcomes', *Psychonomic Bulletin & Review*, 1: 383–389.

Green, L., Myerson, J., Shah, A. K., Estle, S. J., & Holt, D. D. (2007) 'Do Adjusting-Amount and Adjusting-Delay Procedures Produce Equivalent Estimates of Subjective Value In Pigeons?', *Journal of the Experimental Analysis of Behaviour*, 87(3): 337–347.

Hare, C. (2008) 'A Puzzle About Other-Directed Time Bias', *Australasian Journal of Philosophy*, 86(2): 269–277.

Hausman, J.A. (1979) 'Individual Discount Rates and the Purchase and Utilization of Energy-Using Durables', *Bell Journal of Economics*, 10(1): 33–54.

Hill, T. E. (2002) *Human Welfare and Moral Worth*. Oxford University Press.

Holt, D., Green, L., Myerson, J., & Estle, S. J. (2008) 'Preference Reversals With Losses', *Psychonomic Bulletin & Review*, 15: 89–95.

Kirby, K. & Herrnstein, R. J. (1995) 'Preference Reversals Due to Myopic Discounting of Delayed Reward', *Psychological Science*, 6(2): 83–89.

Lloyd, H. (2021) 'Time Discounting, Consistency, and Special Obligations: A Defence of Robust Temporalism', *Global Priorities Working Paper*, 11: 1–38.

Loewenstein, G., & Elster, J. (eds) (1992) *Choice Over Time*. Russell Sage Foundation.

Mazur, J. E. (1987) 'An Adjusting Procedure for Studying Delayed Reinforcement', in M. L. Commons, J. E. Mazur, J. A. Nevin, and H. Rachlin (eds), *Quantitative Analysis of Behaviour V: The Effect of Delay and of Intervening Events on Reinforcement Value*. Erlbaum.

——— (2000) 'Tradeoffs Among Delay, Rate, and Amount of Reinforcement', *Behavioural Processes*, 49(1): 1–10.

Mogensen, A. (2022) 'The only ethical argument for positive δ? Partiality and pure time preference', *Philosophical Studies*, 179(9): 2731–2750.

Noggle, R. (2009) 'Give Till it Hurts? Beneficence, Imperfect Duties, and a Moderate Response to the Aid Question', *Journal of Social Philosophy*, 40(1): 1–16.

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

Plato. (1997) 'Protagoras', in J.M. Cooper (ed), *Plato: Complete Works*. Hackett Publishing Company.

Pummer, T. (2023) *The Rules of Rescue*. Oxford University Press.

Purves, D. (2016) 'The Case for Discounting the Future', *Ethics, Policy, & Environment*, 19(2): 213–230.

Rawls, J. (1971) *A Theory of Justice*. Harvard University Press.

Riedener, S. (2025) 'Authenticity, Meaning, and Alienation: Reasons to Care Less About Far-Future People', in H. Greaves, J. Barrett, & D. Thorstad (eds), *Essays on Longtermism*. Oxford University Press.
Samuelson, P. (1937) 'A Note on Measurement of Utility', *The Review of Economic Studies*, 4(2): 155–161.
de Shalit, A. (2021) 'The Community, the Nation, and Obligations to Future Generations', in S. Gardiner (ed), *The Oxford Handbook of Intergenerational Ethics*. Oxford University Press.
Sidgwick, H. (1907) *The Methods of Ethics*. Macmillan.
Slote, M. (2003) 'Empathy, Immediacy, and Morality', in H. Dyke (ed), *Time and Ethics: Essays at the Intersection*. Springer.
——— (2007) *The Ethics of Care and Empathy*. Routledge.
Stern, N. (2008) 'The Economics of Climate Change', *American Economic Review*, 98(2): 1–37.
Strotz, R. H. (1955) 'Myopia and Inconsistency in Dynamic Utility Maximization', *The Review of Economic Studies*, 23(3): 165–180.
Sullivan, M. (2018) *Time Biases: A Theory of Rational Planning and Personal Persistence*. Oxford University Press.
Sung, L. (2025) 'Time Bias and Altruism', *Mind*, 134(534): 373–396.
Velleman, D. (1991) 'Well-Being and Time', *Pacific Philosophical Quarterly*, 72(1): 48–77.

4.

Partial Aggregation, Charitably Construed

Abstract. In this article I consider what recently popular *partially aggregative* principles would imply for the practice of individual charitable giving. My key suggestion is that partially aggregative principles will systematically generate permissions that individuals not donate towards charitable causes – even extraordinarily effective ones. One potential interpretation of this result is that partially aggregative principles thus provide us with independent traction on the long-discussed problem of moral overdemandingness. I raise two important complications for anyone who would view partially aggregative principles as solving the overdemandingness problem *per se*: a puzzle concerning modal fragility and a dilemma concerning reasons weightings.

1. Introduction

Somehow, as often seems to happen, you find yourself in:

> **Teletransmitter Room.** Jones is being electrocuted by a loose cable in the television transmitter room. You can save him by turning off the power. But turning off the power will cause *n*-many million television watchers to miss out upon a few minutes of mild entertainment. They will each suffer one small 'grump' as a result.[1]

In **Teletransmitter Room**, different persons have claims upon you for aid. First is Jones. And second are the many millions of television watchers. Unfortunately, these claims cannot all be satisfied. What ought you do? Whom ought you help?

[1] Scanlon (1998: 235).

Proponents of *full aggregation* believe that, when an act φ would satisfy the claims for aid presented by separate persons, we should 'stack together' those separate claims such that they achieve a greater weight. Full aggregation seems to imply that you should aid the television watchers. After all, it seems that, for some sufficiently large value of *n*, the collective suffering contained within *n* grumps will outweigh the suffering contained within Jones' death. If that is right, then you ought to spare the television watchers from grumps.[2]

In what is by now a familiar story, many philosophers find the fully aggregative appraisal of **Teletransmitter Room** deeply implausible. They think that one ought to save Jones, no matter how many television watchers are involved. They maintain as much while still granting that, in certain other contexts, it might be appropriate to engage in the aggregation of claims – even if it might be appropriate to spare *n* persons from a grump rather than spare Jones from a lost finger.

I'll refer to such theorists as *partial aggregationists*.[3] I'll gloss the partially aggregative position in the following way: agents ought to maximise the satisfaction of *relevant* claims upon their conduct, where a claim is only relevant so long as it is strong enough compared to those claims with which it competes. When death is on the line for some, the thought goes, the claims of others to be spared from grumps are simply irrelevant. It's this fact which explains why, in **Teletransmitter Room**, you ought to save Jones.

The literature on partial aggregation continues to burgeon.[4] Important work remains to be done, however, in figuring out the exact form and the actual practical implications

[2] One complication is that the badness of harm events might be governed by a lexical ordering. I revisit this complication at the end of §2.
[3] Crucial contributions were Kamm (1993) and Voorhoeve (2014). More recent discussions include Tomlin (2017), Horton (2017; 2018; 2020), Tadros (2019), Walen (2020), Voorhoeve (2020), Mann (2021; 2022), Hart (2022; 2025), Pummer (2023), Mann & Barry (2025), Bourguignon & Mossé (2025), and Williamson (forthcoming).
[4] See fn.3.

of partially aggregative principles. In the following discussion I aid in this task by examining what partially aggregative principles would imply for the practice of individual charitable giving.

The key claim for which I argue goes as follows. In general, partially aggregative rules will generate permissions that individuals *not* donate towards charity – even towards highly effective ones. It turns out that this result holds under various different ways of precisifying partial aggregation to handle risk and uncertainty, and across various charitable contexts.

A natural interpretation of this result is that partially aggregative principles thus have some important role in mitigating against the overdemandingness of our final moral theory: in shielding persons, that is, from a moral requirement that they give every marginal second of their time and every marginal cent of their savings for the sake of very distant others.

Although I find this line of thought intriguing, I stop short of claiming that partially aggregative principles could be understood as furnishing us with a complete solution to the problem of moral overdemandingness *per se*. This is in light of a puzzle and a dilemma. The puzzle is that partially aggregative principles only appear to solve the problem of demandingness in a manner which is exceptionally *modally fragile*. And the dilemma is that, depending upon the reasons weightings employed by the partial aggregationist, it seems that partial aggregationists must choose between a highly demanding all-things-considered moral theory or an all-things-considered moral theory with an extreme level of bias in favour of identified lives.

The argument is structured as follows. In §2, I set out partially aggregative principles, including some of the most intuitive ways of rendering those principles sensitive to risk and uncertainty. In §3–§5, I show how partially aggregative principles will generally

generate permissions *not* to make charitable donations, even highly cost-effective ones. I show, in particular, how this result holds under multiple different interpretations of risk-sensitive partial aggregation and across multiple different kinds of charity. In §6–§7, I interrogate whether partially aggregative principles could thus be used to solve the long-discussed problem of moral overdemandingness. I raise, as I go, the aforementioned puzzle and dilemma facing anyone who would view partially aggregative principles as a solution to overdemandingness concerns *per se*. §8 concludes.

2. Partial Aggregation and Risk

Our first task is to figure out what form partially aggregative principles should take. Partial aggregationists believe, as I've indicated, that we should maximise the satisfaction of *relevant* claims upon our conduct. But unfortunately, there are different ways of interpreting this central maxim when it is applied to contexts of risk and uncertainty.

A familiar distinction can be drawn between the *ex post* and *ex ante* construals of partial aggregation.[5] In extremely broad terms, the *ex post* construal operates by foregrounding structural features of the outcomes themselves which might follow any given act φ. For instance: suppose that φ might lead to outcomes $o_1 - o_n$. We can then deduce the strength of the strongest claim in favour of φ by looking at the strongest single claim in each of the outcomes $o_1 - o_n$ and by taking the probability-weighted sum of these claims.[6] Thus, if φ would either bring about an outcome o_1 in which an individual i_1 is spared death, or an outcome o_2 in which an individual i_2 is spared death, or an outcome o_3 in which an individual i_3 is spared death … and so on … then there would be an extremely strong

[5] N.B. The distinction isn't exhaustive, as in the hybrid models developed by Suikkanen (2019) and Walen (2020).
[6] See, for instance, Horton (2017).

claim in favour of performing φ, equivalent in strength to the claim that a single individual would have to be spared death with certainty. As a result, any minor headaches spared by some alternative act ψ are likely to qualify as deontically irrelevant.

By contrast, the *ex ante* construal operates rather differently. Under the *ex ante* construal, whether or not any given individual *i* has a claim that I φ (rather than ψ) is determined by the *prospects* that *i* would hold under the performance of φ (rather than under ψ). An individual's *prospects*, under φ, are found in the probability-weighted sum of *i's* well-being across each outcome $o_1 - o_n$ which might follow φ's performance.[7] The *ex ante* construal has rather different implications to the *ex post* construal: the *ex ante* construal might judge it perfectly acceptable to implement a policy which is certain to result in *someone* dying, so long as, at the moment of choice, that policy is in the expected interests of every single person $i_1 - i_n$.

The problems facing both *ex post* and *ex ante* partial aggregation are well-known. Nevertheless, many partial aggregationists remain united by the sense that there must exist *some* solution to these problems, and that even seemingly imperfect solutions will have greater plausibility than accepting the verdict that one ought to abandon Jones in **Teletransmitter Room**.[8]

Most famously, *ex post* partial aggregation is vulnerable to the problem of *social risk*.[9] Roughly speaking, many policies improve the expected welfare of every single person within a given reference class even though, in the long run, those policies are certain to bring extremely severe harm to someone or other within the reference class. (Think, here,

[7] We can be more ecumenical: perhaps *i's* prospects could also be identified with the risk-weighted sum of their well-being across $o_1 - o_n$. In what follows I put this complication aside.
[8] A sentiment expressed in Tadros (2019: 203) and Walen (2020: 68).
[9] Seminal discussions are in the context of fully non-aggregative contractualism, as in Ashford (2003) and Frick (2015).

of mass vaccination policies and of airplane travel.) Such socially risky policies often strike us as morally innocuous. But *ex post* partial aggregation will not allow for them. For under those policies, howsoever the world turns out, someone is going to suffer an extreme harm. Under *ex post* partial aggregation, then, the complaints presented by the inevitable 'extreme sufferers' seem to render the benefits reaped by those policies deontically irrelevant, such that nothing can be said in their favour.[10]

The problem of social risk has caused some, including myself, to feel that the *ex ante* construal of partial aggregation has greater plausibility.[11] The *ex ante* model plainly escapes the problem of social risk for the following reason. A socially risky policy such as (say) mass vaccination might be certain to result in someone eventually dying. But nonetheless, that policy may greatly improve the prospects of each individual person i_1 – i_n at the moment of choice. So long as it does so, then no individual person will be able to raise a complaint against that policy. As a result, the policy will qualify as ethically permissible.

Of course, the *ex ante* construal faces its own problems.[12] The most notorious of these is that *ex ante* partial aggregation appears to inappropriately prioritise identified over statistical lives. To see why, consider the following well-known case:

---

[10] For complications, see Horton (2017: 68). In Williamson (forthcoming), I argue that the problems for the *ex post* partially aggregative model are worse than this: there may be no non-extreme claims that the *ex post* partial aggregationist can ever view as relevant, given the way in which seemingly innocuous acts proliferate causal ramifications.

[11] The *locus classicus* is Frick (2015), though, as mentioned, Frick's discussion concerns fully non-aggregative contractualism.

[12] I put aside certain problems for now. In Chapter 6, I confront the 'puzzle of *ex ante* designation', which threatens *ex ante* partial aggregation in addition to various other *ex ante* moral standards. And in Williamson (forthcoming), I nevertheless argue that there is an important class of cases in which the *ex ante* view seems to struggle with the problem of causal ramifications.

> **Mineshaft.** You are an energy magnate who employs 1000 workers in various coalmines. One of your miners, Gareth, is currently facing death. The longwall shaft in which he was drilling has collapsed. You can save Gareth. Alternatively, you could use the money to invest in sweeping safety reforms. By doing so you expect to reduce the number of future collapses and thus prevent the deaths of three (as yet unidentified) miners in the coming year.[13]

When applied to **Mineshaft**, *ex ante* partial aggregation privileges identified others in the following way. Since Gareth is identified at the moment of choice, he has a single (very strong) *ex ante* claim to be spared death. By contrast, those who will be spared death by the safety reforms are not 'identified' at the moment of choice. Each of them thus has a much weaker *ex ante* claim to be given a 3/999 *chance* of being spared death. Thus, *ex ante* partial aggregation risks generating the verdict that you ought to save one (identified) Gareth rather than three (as yet unidentified) miners.[14]

Opinions might differ about whether it is problematic for a moral theory to privilege identified Gareth over three unidentified others: a part of the challenge is that, for better or for worse, an identifiability bias of this kind is deeply ingrained within ordinary moral psychologies.[15] Nonetheless, there comes a point at which *ex ante* partial aggregation's identifiability bias clearly becomes unconscionable. Suppose that there are 100,000 miners in the shaft and that, as before, our choice is between saving identified Gareth versus instituting safety reforms which would save (say) 0.3% of the miners. It is very hard to believe that it would still be permissible to save Gareth rather than 300 unidentified others. But as before, *ex ante* partial aggregation risks implying as much. For Gareth has a far stronger *ex ante* claim to be given rescue than any unidentified miner has to be given the safety reforms.[16]

---

[13] Frick (2015: 216).
[14] For complications, see Horton (2017: 68).
[15] For an overview see Żuradzki (2019).
[16] Frick (2015: 216).

How is the problem to be solved? What seems fairly intuitive, here, is a shift towards *pluralism* about the right-making features which feed into the deontic status of acts.[17] Under pluralism, when determining whether any given act φ is all-things-considered permissible (required), we must ask at least two separate questions:

1. Do the *claims* facts speak in favour of φ? (This is the case insofar as φ is permitted by *ex ante* partial aggregation.)
2. Do the *goodness* facts speak in favour of φ? (This is the case insofar as φ promotes the greater expected goodness.)

It is, in the end, the combination of our answers to questions (1) and (2) that determine whether or not a given act qualifies as all-things-considered permissible. Each criterion endorsed by the pluralist is meant to pick out, presumably, a distinctive way in which the moral facts might speak in favour of a given act. The first criterion, roughly speaking, concerns whether or not a given act is interpersonally justifiable to the particular persons who stand to be affected by your (risk-carrying) choice. The second criterion, roughly speaking, concerns whether or not a given act promotes the good.

If we adopt pluralism of such a kind, we can better cope with **Mineshaft**. We should save Gareth if doing so leaves us with three identified deaths, since the claims facts speak more heavily in favour of saving him. But we should abandon Gareth if doing so leaves us with three hundred identified deaths, since the goodness facts then speak much more heavily in favour of abandoning him. In this way, the pluralist model has at least a certain *prima facie* plausibility in attempting to capture our ordinary judgements about the comparative importance of identified versus statistical lives.[18]

[17] As in Frick (2015: 220), who attributes the suggestion to Parfit (2011: 368–370). The shift to pluralism isn't entirely uncontroversial. Horton (2017: 59–60), for instance, provides a counterexample to the pluralist; it's unclear to me, however, whether the counterexample draws upon intuitions which all *ex ante* theorists are likely to accept.

[18] A potential complication: doesn't this kind of pluralism imply that one might be required to spare 'grumps' in **Teletransmitter Room**, since sufficiently many 'grumps' would cause the goodness facts to

3. *Ex Ante* Partial Aggregation and Charity

That is a brief sketch, then, of some of the most natural ways of extending partial aggregation to cover risk and uncertainty. As foreshadowed, the central and practically relevant question in which I am interested goes as follows: what would partially aggregative principles such as these imply for the practice of individual charitable giving?

I'll argue, as indicated, that partially aggregative principles generally generate permissions that givers *not* make charitable donations, even towards highly effective charities. Such principles might, in this way, have an important role to play in mitigating against the demandingness of our all-things-considered moral theory.

Let me flag that the following discussion is opinionated in the following way. I'll begin by asking what *ex ante* partial aggregation implies about charity, and I'll spend a greater portion of my time on this first question. This is because – cards on the table – my own view is that the *ex ante* interpretation holds greater plausibility. I hope that *ex post* partial aggregationists are willing to bear with me: I eventually suggest that, due to structural considerations about the ways in which many charitable organisations are typically

---

speak to an extreme degree in favour of doing so? This important complication is footnoted in Frick (2015: 223) – unfortunately, I must probably consign it to a footnote too. In short, the most promising solution seems to involve partial aggregationists accepting a *lexical ordering* on their axiology, such that the badness involved in Jones' death can never be outweighed, as an axiological matter, by the badness involved in a collection of n-many persons suffering a 'grump.' This means that the goodness facts, in the original **Teletransmitter Room**, will never speak in favour of sparing 'grumps.' More formally, I find the following sort of model most appealing (from Nebel (2022), with relevant discussion in Portmore (1999), Kamm (2007: 33–37), Thomas (2018), Lazar & Lee Stronach (2019), and Hawthorne, Isaacs, & Littlejohn (2023)). Consider a given distribution of bads X. Suppose that X contains some *important* bads (like electrocution) and some *minor* bads (like grumps). The important bads are those which hold lexical priority over the minor bads. The sum of both the important bads and the minor bads within X is given by the vector (i_X, m_X). We then say that there is some quantity of the important bads which is, itself, capable of outweighing any quantity of the minor bads. Using Δ to represent the quantity of the important bads i such that those bads are capable of outweighing any quantity of the minor bads m, we say that, for any two distributions of the good X and Y, X has at least as great a sum of bad as Y when (1) $i_X - i_Y > \Delta$, or when (2) $i_X \geq i_Y$ and $m_X \geq m_Y$. One might worry that this move towards a lexical axiology makes it redundant to have endorsed deontic partial aggregation concerning *claims* in the first place. This isn't obvious, however. A population $(p_1, p_2, p_3 \ldots p_n)$ might be faced with a choice between two potential distributions of 'important' harm, $(100i, 0, 0, 0, 0, \ldots 0)$ and $(0, i, i, i, i, \ldots i)$. Even if the two distributions contain the same total quantity of important harm i, one might still feel the pull of the partially aggregative sentiment that the second distribution is preferable, at least in part because there is an exceptionally strong single claim in its favour.

organised, similar results hold for *ex post* partial aggregation across a wide range of practically relevant contexts too.

3.1 Statistical Charity

There are, of course, various ways in which well-motivated individuals can donate towards charity. But one broad avenue involves donating towards what we might label *statistical charities*: that is, charities which benefit particular persons, but where the identities of potential beneficiaries cannot be known to you (or perhaps to anyone) at the moment of choice. Various statistical charities are, in short, incredibly effective at saving lives. According to the charity evaluator GiveWell, for instance, one of the most cost-effective means of preventing statistical deaths involves donating money to the Against Malaria Foundation (AMF) in order to fund the distribution of insecticide-treated bed nets in high-risk countries. GiveWell estimates that, in expectation, each additional donation of $3000-$8000 prevents an additional death. For ease of discussion, let's set the number at $5000.[19]

Suppose that you are deciding whether or not to donate to a highly effective statistical charity such as AMF. Does *ex ante* partial aggregation speak in favour of a requirement that you donate? Plausibly, the answer to this question is *no*. Depending upon the details, partial aggregation may even speak in favour of a requirement that you *not* donate.

The reasoning goes as follows. There are different persons who stand to gain and lose in your choice. First of all, of course, there is *you*. You are an identified individual. And paying $5000 is certain to come at a substantial cost *to you*.[20] By contrast, when you donate

[19] See https://www.givewell.org/charities/amf.

[20] Momentarily, I'll revisit the (potentially controversial) assumption that financial giving is prudentially costly. In Chapters 1 and 2, for instance, I suggested that charitable giving is often *not* personally costly. My suggestion in §5 of this Chapter will be that, even if a marginal charitable donation of (say) $5000

to a statistical charity like AMF, there is no single identified individual who is certain to be the beneficiary of your choice. Given that charities such as AMF advertise their recent net distributions, you might be able to form some level of confidence about the particular country (or set of countries) towards which your donation might go. And some time after donating, you might receive an update as to which particular geographic region ended up receiving the nets matched to your donation. But these facts aren't sufficient to undercut the prior claim that, at the moment of choice, there are many millions of persons who stand to be the individual who is aided by your giving.

It is especially worth bearing in mind that, plausibly, in the case of AMF and various other statistical charities, one's donation results in a life's being saved as a result of improving the objective chances for many persons (by some small degree) that they find themselves spared from severe suffering. This is because the donation in question funds the distribution of many nets, each of which (in turn) reduces an individual's chances of suffering harm by some small margin. If this is right, there may in fact be *no* person who is capable of identifying (*ex ante*) the particular individual who your donation would end up saving.

Under *ex ante* partial aggregation, then, the correct characterisation of your decision appears to go as follows. On the one hand you have a weighty claim against the act of donating – the $5000 causes you substantial loss with certainty. On the other hand there are very many individually small claims upon you to donate – claims to be given, say, a 1/10,000,000 chance of being *the one* who is spared suffering.[21] But it is exceedingly plausible that your own complaint against being caused a $5000 loss with *certainty* would

towards AMF were costless to me, it is unlikely that it would be costless to my 'nearby surrogates' in a sense which is currently of relevance for the partial aggregationist.

[21] The assumption, here, is that you can broadly estimate a particular country of 10,000,000 persons towards which your nets will be distributed.

outweigh any individual claim offered by a stranger to be given a 1/10,000,000 chance of being spared sickness or death. And it's plausible that this complaint would outweigh each such individual complaint by a margin capable of generating irrelevance. If this is right, then *ex ante* partial aggregation leans towards a permission that you not donate (rather than a requirement that you do).[22]

If this result holds with respect to AMF, then, as a rule of thumb, it will hold with respect to many other statistical charities in general. This is simply because AMF is one of the most cost-effective methods via which affluent persons can prevent grave statistical harms. That is: donating to AMF is one of the ways in which altruists can prevent a grave statistical harm such as death while suffering the smallest possible burdens themselves.

3.2 Identifiable Charity

By contrast, many kinds of charitable donation seem to benefit persons who are *identifiable* at the moment of choice. One might suggest that *ex ante* partially aggregative principles nevertheless require such donations. After all, as we've already seen, *ex ante* partial aggregation generally exhibits a bias in favour of identified rather than statistical lives. I think, however, that this tempting suggestion is misleading – at the very least, it is misleading for a wide range of practically relevant cases.

It goes without saying, of course, that there are many different ways in which individuals can engage in identifiable charity from a distance. One salient case type, which

[22] Now, it certainly seems that *fully non-aggregationist* principles couldn't generate a requirement to donate, since, plausibly, a claim to be spared a guaranteed loss of $5000 is greater in strength than any single claim to be spared a 1/10,000,000 chance of death. But still, one might wonder if the difference between these claims isn't great enough for the stronger claim to render the weaker ones irrelevant. This is an important possibility. My own intuition is that the difference could indeed be large enough. Suppose that a sudden unexpected loss of $5000 was great enough to undermine a substantive and important one of my personal projects. Someone then tells me that I should give up my project in order to improve their chances of survival of a stranger by 1/10,000,000. I generally have the sense, here, that the stranger could not press their claim upon me in a sense which was indicative of relevance, and that this fact would be attributable to the difference in claim strength.

we can put aside in what follows, is to donate from a distance towards emergency rescue scenarios. Broadly, these are cases in which some small number of identified persons have gone missing or entered a state of peril and in which costly efforts can be made to rescue them. (A boat has become lost at sea, perhaps, or a hiker has been missing in the woods for three weeks.) For at least two reasons, we can put these cases aside in what follows. First, it is often not obvious whether individuals *are* routinely positioned to help via donation in such cases: in at least many such cases, aid is systematically administered as a matter of course by state bodies, government bodies, or emergency services. But second, suppose that you can indeed help via donation in such a case (perhaps by donating towards a rescue fund): it is well-known that it is extremely difficult to justify emergency rescues such as these even on grounds of total utility, given the beneficiaries involved are so few and the costs involved in enacting any single rescue are so astronomically high. In light of this fact *ex ante* partial aggregation – like various other normative standards – will struggle to generate a requirement that one donate.[23]

Still, there are certain other kinds of identifiable charity, where *ex ante* partially aggregative principles seem like they might indeed require one to donate towards those charities. I particularly have in mind the following example. What we might call 'letter writing charities' allow charitable givers to select an individual person facing harmful conditions, even from a very great distance, and commit to a direct and ongoing charitable

[23] For recent discussion of the puzzle of costly rescue, see Sinclair (2021). Now, often one does indeed have the opportunity to donate towards charities which enact emergency rescues, such as, for instance, the Scottish Mountain Rescue. However, this is a point at which further empirical complications must be kept in mind. Whether or not a rescue service sends out (say) a particular helicopter on a particular mission is not usually contingent upon whether or not the rescue service receives funds in an appeal at the time of *that mission in particular*. This is because the rescue service in question is operated by volunteer labour or by using donations of materials and transport. Donations instead facilitate institutional structures which allow for the continued support of future missions. This, of course, does not mean that the donations in question are any less vital. But it does mean that the donation implicitly secures benefits for (as yet) unidentified victims. So, for many of the ways in which we might donate towards 'identified' charities, those donations in fact seem to save the lives of as yet unidentified persons.

commitment *to that person*, such that *they in particular* are spared from harmful conditions. To provide one such example: by visiting the website of World Vision or similar programs, one can find a list of identified persons living in conditions of poverty whom one can choose to sponsor.[24] Choosing to sponsor a child commits one to the long-term provision of aid for that particular child, often characterised not just by financial donations but also by long-term written correspondence, the formation of a personal connection, and so on.

It's extremely tempting to think that, when it comes to letter writing charities, *ex ante* partial aggregation would indeed generate a requirement to provide aid. After all, these are situations in which I become aware of a particular identified individual – call her Alice – who faces great suffering, and whom I can directly aid by making a much smaller sacrifice on a regular basis.

What I want to suggest, however, is that even this conclusion may be too quick. Upon closer inspection, *ex ante* partial aggregation may not even support a requirement to donate towards letter writing charities. Why? Well, for a start, in order to know whether *ex ante* partial aggregation requires that you donate towards an identified individual like Alice, we need to know some important further details. One needed detail is whether there are reasonable chances that Alice would have been funded by someone else under the counterfactual baseline according to which you *hadn't* donated.

Many persons read the homepage of the World Vision website and make a subsequent decision to donate. It may thus be plausible that, if *you* read the webpage and decide to donate to a particular person Alice, then your donation *with respect to Alice* will be overdetermined: had you not donated *to Alice in particular*, then somebody else would have done so. Of course, whether or not such systems of charitable aid do contain

[24] See, for instance, https://www.wvi.org/.

overdetermination is a difficult matter to settle, and seems to depend, at least in part, upon facts to which you will not have access in making your decision (such as the algorithm via which any given individual is selected to be displayed upon the website homepage). But there is at least some evidence in favour of the overdetermination hypothesis: it is a standard feature of such websites that donors can opt-in to fund the individual who has been waitlisted for aid the longest.[25] If you opt-in in such a way, then under the assumption that at least some other donors use the same filter, your donation will be subject to overdetermination. It will be the case that, had *you* not aided the longest-waiting identified individual, then somebody else would have done so.

The possibility of overdetermination greatly complicates the final analysis of things, here. One possibility, for instance, is that donations towards letter-writing charities effect change via the realisation of *tipping points*.[26] That is, there is a high probability that my individual donation makes no counterfactual difference to the welfare of other persons – everyone on the list, say, will be funded regardless of whatever I do. But there is some very small chance that my decision will make a very large difference for very many (as yet unidentified) persons – that is, there is a small chance that my act constitutes a tipping point by indicating to program administrators that they have sufficient interest in order to expand to a new community, or to increase the number of beneficiaries for which they advertise for in any given calendar year.[27]

Insofar as one's donations towards letter-writing charities effect change via a tipping point structure, then it becomes far less likely that *ex ante* partial aggregation generates a requirement that I donate. For if that is how these cases are best characterised, then there

[25] See, for instance, https://www.worldvision.org.uk/sponsor-a-child/.
[26] For related discussion, see Cullity (2004: 58–60), who considers the possibility that typical donations to famine relief are subject to a tipping point or threshold structure, such that there is in fact a very large chance that my donation to famine makes no difference, but some small chance that it triggers (say) an institutional decision such that an additional food consignment is ordered.
[27] For canonical discussion of such cases, see Singer (1980), Norcross (2004), and Kagan (2011).

are in fact very many (unidentified) persons who each have very weak *ex ante* claims upon me to be given a *very small chance* of being aided (by my provision of a donation which turns out to constitute a tipping point, such that they are provided aid when they would not have otherwise been aided). But making the donation nevertheless comes at a certain cost to myself. So again, my own claim against donating is likely to render the much smaller *ex ante* claims presented by others irrelevant.[28]

4. *Ex Post* Partial Aggregation and Charity

I'm eager to deal with objections. But I've so far skipped over *ex post* partial aggregation. And *ex post* partial aggregationists might reasonably ask: what about me? Let me briefly explain why *ex post* partial aggregation seems likely to entail the same result across a large class of practically relevant cases.

As it so happens, we've already stumbled across the needed resources. We've already seen that many international charities, whether targeted towards identified or non-identified beneficiaries, seem to effect change via a tipping point structure.[29] That is: if I decide to donate to Charity X, then there is a very large chance that my donation will make no counterfactual difference at all to the welfare of others. But if I decide to donate to Charity X, there is a very small chance that my act may make a very large difference to the welfare of *very many* persons (by allowing Charity X to expand into some additional region, say, or allowing them to order an additional large consignment of aid).

---

[28] A separate and more mundane point is that very few persons have actually visited the homepage of websites such as World Vision. For the large majority of persons who have *not* visited the webpage of World Vision, then the individuals who might be listed there and in need of aid are nevertheless *ex ante* unidentified. One might suggest that, prior to action, agents must somehow try to 'deanonymise' as many potential beneficiaries as possible. But it's hard to deny that this move introduces complications too: each of us could, for instance, spend our entire lives attempting to deanonymise as many potential beneficiaries as possible.

[29] As above, see Cullity's (2004: 58–60) extended discussion of famine relief.

Insofar as tipping points regularly characterise the structure of charitable organisations, then *ex post* partial aggregation will struggle to deliver the verdict that one ought to donate towards those organisations. How so? Well, consider the act of donating $5000 to a tipping point charity. Whatever happens, there is *someone* who has a moderately strong claim against donating – namely, myself, who will be burdened in each potential outcome which might follow my act.[30] This suggests that there is quite a strong *ex post* claim against donating.[31] And it is only in one (extremely unlikely) outcome o_n that *anyone* has a claim upon me to donate towards those charities – namely, the (extremely unlikely) outcome in which my donation serves as a 'tipping point' for the promotion of a great deal of good. This suggests that there are only a weak set of *ex post* claims in favour of donating. If the difference in strength between these two claims is sizeable enough, then the *ex post* claim *against* donating will render each of the *ex post* claims in favour of donating irrelevant. And we will again receive the same verdict that it is ethically acceptable to refrain from donating.[32]

This point isn't entirely novel. A similar point is recently made by Emma Curran with respect to altruistic donations aimed towards the reduction of existential risk.[33] Curran notes that it is difficult for *ex post* partial or *ex post* non-aggregationists to justify existential-risk reducing interventions when an alternative on offer would (say) involve benefitting some comparatively small number of present persons with certainty. After all, upon reflection, there are only extremely weak *ex post* claims in favour of an 'existential

[30] As before, this assumes that donating marginal funds is costly, which I revisit shortly.
[31] As above, we find the strength of the strongest *ex post* claim in favour of φ by taking the probability-weighted sum of the strongest claim featuring in o_1, the strongest claim featuring in o_2 … and so on.
[32] For instance, suppose that my individual donation has a 0.0001 chance of being a trigger of serious aid for very many persons. The strongest claim in favour of donating would then be notably weak, consisting of a 0.9999(claim of nothing)+0.0001(claim to be given serious aid). The second strongest claim in favour of donating would be exactly as weak, consisting of a 0.9999(claim of nothing)+ 0.0001(claim to be given serious aid) … and so on.
[33] Curran (2025).

risk' donation which carries (say) a 1/1,000,000,000 chance of saving 100,000,000,000 future lives but which would otherwise do nothing. By contrast, there appear to be much stronger *ex post* claims in favour of a donation directed towards present persons which substantially improves the welfare of 1000 people with certainty. What Curran does not quite turn to consider, however, is that analogous reasoning could be extended towards a case involving only present persons. Many altruistic interventions directed towards *present* beneficiaries appear to be governed by a tipping point structure, such that they carry a very small chance of 'pushing things over the line' and providing additional aid to very many persons. If this is right, then the *ex post* claims in favour of donating towards those charities seem quite weak. *Ex post* anti-aggregative views will generally struggle to generate a requirement in favour of making those donations – at least insofar as those donations come with a certain cost to the giver herself.

A great deal depends upon the details here. *Ex post* partial aggregationists might deny, for instance, that certain kinds of charitable donation generate change via tipping points – donations to famine relief, say. I don't claim that they couldn't. But at least one alternative analysis of such cases, for those who reject tipping point analyses, is that my individual donation towards famine relief just *can't* make a perceptible difference in such cases, since my donation is utterly minuscule within the grand scheme of international supply chains. If this is right, it becomes even harder for *ex post* partial aggregation to generate a requirement to donate: it is unclear whether there is *any* claim for me to donate if my donation makes non-perceptible differences to the welfare of others.[34]

[34] Donations might make, for instance, imperceptible increases to the quantity of food received by very many separate persons. Something like the 'imperceptible difference' view is ultimately preferred by Cullity (2004: 60). See relevant discussion in Parfit (1984: 76), Nefsky (2011), and Budolfson (2019).

5. Objections

I've explained why, for a very wide variety of charitable causes, partial aggregative principles generate permissions that we *don't* donate. I've skipped over some vital objections, though. Before turning to interpretative questions, let me address these objections in full.

Objection 1: Claims upon Oneself

Central to my above discussion was a potentially controversial assumption. This was that the *agent herself* can have a claim against φ-ing, in virtue of the costs to herself involved in φ-ing, and that this claim is the sort of thing capable of rendering the claims presented by other persons irrelevant. This assumption allowed us to prove that, under partially aggregative principles, agents generally have a permission not to donate towards charitable causes. One might want to reject this theoretical assumption and thus reject my practical conclusions.

I want to provide a conciliatory response to this objection. One thing I can certainly agree with is that there are many institutional decision-making contexts in which we might want partially aggregative rules to be operative and in which it may not even *make sense* to talk of such things as 'claims presented by the decision makers themselves.' For instance, as discussed at length by Alex Voorhoeve, there is an increasing interest in partially aggregative rules as a guide for medical bodies to the allocation of healthcare resources.[35] In such contexts we are not considering, *ex hypothesi*, the personal costs facing the decision-makers themselves – the personal costs facing the assembled chairs of the board, say.

[35] Voorhoeve (2020).

Nevertheless, when we turn to cases of individual decision making, it strikes me as perfectly reasonable to speak of the agent herself as having a 'claim' in favour of some choices rather than others, where the agent's claim has the power to render the weaker claims presented by others irrelevant (just as the claims presented by *others* have the power to render the weaker claims presented by others irrelevant). Consider, for instance:

> **Teletransmitter II**. Jones is being electrocuted by a loose cable in the television transmitter room. You may save him by turning off the power. But turning off the power will cause *n*-many million television watchers to miss out upon a few minutes of mild entertainment. They will each suffer one small 'grump' as a result. Unfortunately, you are Jones.

Insofar as we have even remotely anti-aggregationist sympathies, it seems that we will want to evaluate **Teletransmitter II** in the following way: I, Jones, have some kind of 'claim' to be spared from electrocution. Furthermore, this claim is so great in weight that it outweighs the claims of the television watchers to be spared a grump, no matter how great their number. This is true even if, perhaps, I would be required to sacrifice my life to save *n*-million persons from crippling paraplegia. In light of cases like **Teletransmitter II**, I take the above assumption that the agent can present her own 'claim' to be spared from loss or harm, in these contexts, to be relatively uncontroversial.

Some, of course, may baulk at the idea that persons could hold a literal *claim* upon themselves.[36] My own sense is that it is unimportant whether we think that there are literally such things as 'claims upon oneself' in the above cases, or merely some other moral phenomenon capable of playing the functional role of a 'claims upon oneself' – a *relevant personal interest*, say, which plays the same functional role of rendering the much

[36] Presumably, the question is intimately related to whether or not there can be self-directed duties. For just one model of how self-directed duties might work, see Hills (2003).

weaker claims presented by others deontically irrelevant in settling the question of how you ought to act.

Objection 2: Giving Isn't Costly

In the above discussion, I didn't merely assume that costs to the agent can render the claims of others irrelevant. I also assumed that giving substantial financial donations towards charity is *costly* – costly in the particular sense that it makes agents substantially worse-off than they otherwise would have been. One might reject this assumption and thus my practical conclusions.

I take this objection seriously. I have argued elsewhere, for instance, that charitable giving is often *not* costly in some of the ways which philosophers commonly presuppose.[37] What philosophers particularly tend to neglect is that there is a large body of empirical research on the welfarist benefits associated with regularly making charitable donations.[38] Amongst other facts, for instance, functional magnetic resonance imaging suggests that financial giving serves as a pleasure stimulator; more than this, it is regular prosocial spending (rather than regular personal spending upon oneself) which tends to correlate with higher self-reported happiness.[39]

There are a couple of points to make in response to this concern. The first mundane point is that, since a single donation to a charity like (say) AMF will ultimately benefit a select few persons out of a reference class containing many millions, the certain costs to me involved in donating may only need to be quite small in order to render the *ex ante* claims of others irrelevant, given the apparatus of *ex ante* partial aggregation. If an

[37] See Chapter 1. Admittedly, much of my focus in Chapter 1 is on the seeming costlessness of making charitable precommitments at the start of one's moral life; things could change in relevant ways, perhaps, when we turn to consider agents who are *not* at the start of their moral lives.
[38] For an overview see MacAskill, Mogensen, & Ord (2018)
[39] Harbaugh, Mayr & Burghart (2007) and Dunn et al. (2008).

unexpected donation of $5000 to AMF would merely harm me to *some* degree – perhaps by undermining one of this year's personal projects to which I am passionately committed – then the aforementioned practical conclusions may still hold.

But the second point is that the current objection can be overcome by making only a small structural change to the original argument. For the vast majority of us, our lives are characterised by existing special relations with family, friends, and close acquaintances – call these our 'nearby surrogates.' Whenever I contemplate a donation to a global charity, it's usually an alternative option within my purview to instead use the money to benefit one of my 'nearby surrogates.' Even comparatively small benefits to nearby surrogates are likely to render the expected claims of those benefitted via statistical or identified charities irrelevant, since benefits to nearby surrogates obtain with something much closer to certainty. Furthermore, insofar as donating to global charity typically *benefits me* (by providing me with a sense of satisfaction, meaning, or purpose), similar kinds of benefits accrue *to me* when I use my money to benefit nearby surrogates.[40] The possibility of pursuing benefits for nearby surrogates, then, can still generate the result that partially aggregative standards generally speak against a requirement to donate.

6. Partial Aggregation and Demandingness

Let's take stock. I've argued that partially aggregative principles will not generally support a requirement to engage in substantial levels of charitable giving, even when it comes to extremely cost effective charities such as AMF. In each case, the kernel of insight which allowed us to draw this conclusion was that charitable giving usually comes with certain costs to the identified giver herself (or to her identified nearby surrogates).

[40] Indeed, Dunn et al. (2008) combine direct interpersonal gifts and financial charitable giving into a single index of prosocial behaviour, which turns out to accrue the relevant prudential benefits.

Sometimes, it seems, partially aggregative principles may even do more than this. They may *actively forbid* giving towards effective charitable foundations. Perhaps I am still permitted to donate towards AMF if doing so comes at a certain cost to me, since I can, presumably, waive my own 'claim' against donating. But suppose instead that my elderly neighbour has broken her arm (a $5000 fix) and that nobody else is positioned to provide her with aid. In this case, partially aggregative principles may *actively forbid* that I abandon her, forbidding me from (say) channelling my $5000 towards an effective international charity such as AMF. In such a case, the person with a markedly strong (and seemingly uniquely relevant) claim for aid appears to be my elderly neighbour.

There are clearly different ways of taking these practical results. One possibility, for instance, would be to take these practical results as a *reductio* of partially aggregative principles. Surely we *must* give large portions of our disposable resources towards effective international charities; we must reject any moral principle which tries to deny as much.[41]

Nonetheless, many philosophers may be tempted to view things in a more optimistic light. They may instead conclude that partially aggregative principles have an important role to play in mitigating against the *overdemandingness concern*: that is, in mitigating against a conception of morality which imposes demands which are, in some intuitive sense, too burdensome to bear.[42] Such principles help us resist the sort of requirement which Garrett Cullity once labelled:

> *The Severe Demand.* I should continue to contribute to aid agencies increments of time and money each of which is large enough to allow those agencies to save a life, until either: (a) there are no longer any lives to be easily saved by those agencies, or (b) contributing another increment would itself harm me enough to excuse my failing to save any single life directly at that cost.[43]

[41] Perhaps following Goodin's (2009) sentiment that moral norms must not be underdemanding.
[42] For some of the crucial interrogations see Kagan (1989), Murphy (1993; 2000), Hooker (2000), Mulgan (2001), Ashford (2003), Cullity (2003; 2004), Sobel (2007), Noggle (2009), Timmerman (2015), Chappell (2019), Thomson (2022), Pummer (2023).
[43] Cullity (2003: 403).

For such philosophers, the fact that partially aggregative principles systematically lean towards a permission to *not* donate towards international charity is not to be viewed as a vice of those principles, but rather, as a virtue. Such principles vindicate our ordinary sense that, though I might be required to help my elderly neighbour with her broken arm, I need not sacrifice every cent of my bank account and every second of my time towards effective international charities.[44]

Which of these interpretations is correct? To some extent, theorists' intuitions may lean in opposite directions, and it may be impossible to try and disabuse theorists of those intuitions in fully theory-neutral terms. My goal, in what follows, is more circumscribed. It is to raise a puzzle and a dilemma for anyone inclined to view partially aggregative principles as a complete solution to the overdemandingness concern. At the very most we would be warranted in claiming – as I have done – that such principles may facilitate a certain kind of progress on the problem.

7. A Puzzle and a Dilemma

7.1. The Puzzle

Let's begin with the puzzle. The puzzle, in brief, is that partially aggregative principles only seem to solve the problem of moral overdemandingness in a manner which is exceptionally *modally fragile*.

Now, amongst those who take the demandingness concern seriously, one possibility is that we are after principles of beneficence which impose intuitively reasonable levels of

[44] This interpretation is not entirely new; for instance, similar sentiments were explored (and ultimately rejected) in Unger's (1996: 58) discussion of statistical lives.

demand in the *actual* world. Thus, the reason that (say) act utilitarianism is an objectionably demanding principle is that the principle imposes overly strenuous demands in the world as it is.

As a historical matter, however, many have found this attempted characterisation of the demandingness concern inadequate. Many have felt that what should matter, in evaluating theory demandingness, is something much closer to whether or not a theory imposes reasonable demands across an *appropriate range* of possible worlds.[45]

Of course, the devil is in the details. It has proved notoriously difficult to specify just what this 'appropriate range' might amount to. One prominent view has been that *emergency worlds* – that is, worlds filled with natural disasters and communicable diseases – fail to qualify as appropriate in the relevant sense.[46] The general idea, here, is that emergency worlds have clearly 'broken down', and that we should not be too surprised if the correct moral rules turn out to impose stringent demands in those worlds which have broken down. A natural thought is that this framework might vindicate (say) act utilitarianism as an acceptably demanding theory, for although act utilitarianism is highly demanding in the actual world, the actual world turns out (upon reflection) to be one of these 'broken down' worlds. This line of thought remains controversial, however. For there nonetheless seem to be various worlds lacking emergency features in which utilitarianism would impose extremely strenuous demands.[47]

We need not pinpoint the exact conditions which demarcate the appropriate set of worlds in order to see the relevance of this complication for the current discussion. Partially aggregative principles might, as I have observed, mitigate against the

[45] For relevant discussion see Scheffler (1992), Murphy (2000), Ashford (2003), Hills (2010), and Mogensen (2021). In my view, beyond these discussions, which sometimes treat the problem in passing, philosophers have not yet paid adequate attention to the difficult question of whether modal facts are relevant in evaluating considerations of demandingness. A recent exception is York (2024).
[46] A position associated with Scheffler (1992: 100).
[47] See Hills (2010) and Mogensen (2021).

demandingness of morality in the actual world. But partially aggregative principles nevertheless seem to impose extremely strenuous demands in nearby and morally relevant states of affairs. Insofar as that is the case, then partially aggregative principles may only constitute a very limited kind of progress on the demandingness objection.

To see why, imagine a nearby possible world in which human society followed the same broad historical trajectory as our own – Alexander conquered widely and his empire vanished quickly, the Napoleonic forces were defeated at Waterloo, the Berlin Wall eventually fell, and so on. It's a quirk of this world, however, that contemporary charitable institutions like World Vision operate rather differently. Relying on guilt as a powerful motivator, these charities go to affluent countries and assign each potential donor a number (1, 2, 3…). They then go to geographic areas containing extreme poverty and assign each potential beneficiary a number (1, 2, 3 …). Since there are more potential beneficiaries, their numbers repeat. Since there are fewer donors, their numbers do not repeat. If I *qua* donor receive the number 437, for instance, this means that my donations are uniquely twinned to the various no.437 beneficiaries, whose names (we can suppose) are revealed to me. The charities in question tell me that, if I do not provide aid to the no.437 beneficiaries, then nobody else will: those beneficiaries will be utterly abandoned. Say, of any such world, that it contains a system of *Twinned Philanthropy*.

The important point to emphasise here is that, in any world containing Twinned Philanthropy, partially aggregative standards appear to impose extremely stringent demands upon affluent persons, perhaps demands of a kind with those issued by act utilitarianism in the actual world. Each altruist, in such a world, finds herself uniquely and unilaterally positioned to aid many persons, where the certain claims of each such (identified) person will be relevant and weighty, even though the cumulative cost of answering each such claim will be extremely weighty for the altruist herself. So, partial

aggregation ends up imposing extreme demands in worlds much like ours – indeed, in worlds containing social arrangements which could have come about in ours.

Perhaps the response, here, could be that partially aggregative standards still count as making progress on the problem of demandingness, for they nevertheless turn out to impose strenuous demands within *far fewer* relevant states of affairs than does (say) standard act utilitarianism. But as Mogensen notes in a slightly different context, it remains unclear what this suggestion amounts to.[48] If, for instance, we are concerning ourselves with the tapestry of possible worlds, then there are going to be an infinite number of possible worlds containing Twinned Philanthropy and in which partially aggregative standards would nevertheless impose extreme demands.

Perhaps the reply could be that Twinned Philanthropy worlds do not, in the end, count as falling within the *appropriate range*. This avenue might be more promising. But as far as my own intuitions go, I am not so sure. Unlike Scheffler, I am tempted by the idea that the actual world *actually is* within the appropriate range of worlds, and that this fact best explains why so many moral theorists have attempted to respond to the accusation of demandingness by showing that their theory is not demanding in the world as it actually is.[49] And it is not obvious to me that Twinned Philanthropy worlds differ from the actual world in those respects which would be relevant for assessing theory demandingness: after all, Twinned Philanthropy worlds very closely resemble our own. But if the actual world is in the appropriate range, and if Twinned Philanthropy worlds do not differ from the actual world in the respects which would be relevant for assessing theory

[48] Mogensen (2021: 577).
[49] Scheffler (1992). Famously, for instance, Jackson (1991) responds to the alleged demandingness of consequentialism by showing that, given the structural features underpinning human society as it is or would be under a broad range of variance, consequentialist principles generally allow for the display of special concern towards one's nearest and dearest.

demandingness, it seems that Twinned Philanthropy worlds would indeed lie within the 'appropriate range' after all.

7.2. The Dilemma

Let's put the above puzzle concerning modal fragility aside – let's suppose, for instance, that a modally fragile solution to the problem of overdemandingness is adequate. Nevertheless, it still seems to me that we are left with a difficult dilemma. The dilemma has the following shape: partial aggregationists must either accept a *highly* demanding all-things-considered account of our moral obligations, or accept a particularly vicious version of the identifiability bias. I'll illustrate the dilemma, here, via reference to the *ex ante* partially aggregative model.[50]

Here is how the dilemma works. We already saw, in §2, that *ex ante* partially aggregative principles will need to be embedded within a pluralist framework to retain plausibility. When asking whether a given act φ is required, we must ask whether φ succeeds with respect to the 'claims-based facts.' And we must also ask whether φ succeeds with respect to the 'goodness-based facts.' Somehow, it will be the balance of these two facts which together determine whether or not φ is all-things-considered morally required. But of course, if this is right, then we face a difficult decision. We must decide the comparative weight to assign to the 'claims facts' versus the 'goodness facts' when settling the deontic status of any given act.

As a starting point we might hope to make use of (roughly) the weightings we originally intuited back in **Mineshaft**. That is: the 'claims-based reasons' you have for

[50] As we've seen, it's less clear whether *ex post* partial aggregation *always* makes progress on problem of demandingness. In focussing on the *ex ante* version, then, I restrict my attention to what we might reasonably view as the most promising potential 'demandingness solution.'

saving an identified someone (Gareth) from death are similar in weight to the 'goodness-based' reasons you have for preventing three additional deaths.

If this is our strategy, though, then it nonetheless seems likely that we will still arrive at an exceptionally demanding all-things-considered account of morality. This is because certain altruistic interventions appear to succeed to an *extreme* degree with respect to the goodness-based facts, even if it is unclear, in the end, how those interventions fare with respect to the claims-based facts.

Two well-discussed examples come to mind. First, as recent literature emphasises, certain kinds of longtermist interventions directed towards the flourishing of distant-future generations appear to be associated with astronomical quantities of expected good, given the potential span of time over which human life might flourish. A famous estimate from Nick Bostrom estimates, for instance, that we should be indifferent between the expected goodness involved in saving 100,000,000 human lives with certainty versus reducing the chances of premature extinction by a millionth of a percent.[51] Of course, these numbers are controversial, and a great deal has recently been written about whether they are much too grand.[52] Suppose that these numbers *are* much too grand, by several orders of magnitude. It would nevertheless seem that we are left with extremely weighty 'goodness-based reasons' for pursuing longtermist altruistic interventions, interventions such as existential threat reduction. This would hold even if there were only very weak 'claims-based' reasons for pursuing those interventions – even if we felt that it was somehow unclear whether potential future persons hold 'claims' upon us in the sense in which the partial aggregationist is interested.[53]

[51] Bostrom (2013).
[52] For recent complications, see MacAskill (2022), Thorstad (2023), and Mogensen (2024).
[53] Perhaps in light of the non-identity problem, as in Parfit (1984). For recent discussion of non-aggregationism in the context of longtermism, see Curran (2025).

Second, as another contribution to recent literature emphasises, interventions directed towards the reduction of wild animal suffering often appear to be associated with astronomical quantities of good. According to an extremely rough estimate from O'Brien, for instance, we should minimally expect 18.2 million years of extreme wild animal suffering to come to pass within any given calendar year.[54] Many interventions, including mundane ones such as culling, and more speculative ones such as the introduction of edited genes into wild populations, hold the potential to reduce this number by considerable margins. Such interventions appear to succeed to an extreme degree with respect to the 'goodness-based reasons' even if it strikes us as unclear how they fare with respect to the 'claims-based reasons' – even if it struck us as somehow unusual to speak of animals as holding 'claims' upon us in the sense in which the partial aggregationist is interested.[55]

If, then, we endorse the sorts of reasons weightings initially gestured at in **Mineshaft**, we nevertheless seem to arrive at a highly demanding all-things-considered picture of our moral obligations. The fact that the 'claims-based facts' do not generally speak in favour of making charitable donations will neither be here nor there, for certain classes of charitable intervention will be supported (to an extreme degree) by the 'goodness-based facts': the goodness facts will speak in favour of an all-things-considered requirement to donate towards those interventions.

Of course, the above model is just one possible account of how the claims reasons might weigh against the goodness reasons. We could resist a strenuously demanding all-things

[54] O'Brien (2025: 349).

[55] See O'Brien (2023: 16–19) for a recent survey of many proposed interventions on wild animal welfare, and Johannsen (2017) on the particular case of gene editing. Many such proposals are risky, highly speculative, or currently technologically impossible, but as O'Brien notes, given the potentially astronomical stakes involved, this doesn't seem to exempt us from a requirement to pursue those interventions: rather, it seems, rather, to shackle us with a requirement that we devote our resources towards the proper investigation or development of those interventions. For discussion of less speculative interventions such as culling, see Hills (2010).

considered account of morality by simply saying that the claims reasons have *far greater* comparative weight than the goodness reasons when it comes to determining deontic status. If this were the case then there might not be a strenuous moral imperative to donate the vast majority of one's time and resources towards (say) longtermist or animal welfare interventions: such a requirement might succeed (to an extreme degree) with respect to the goodness reasons but it would fail with respect to the claims reasons (which are themselves weighted to an extreme degree).

Unfortunately, this manoeuvre also leads to trouble. Recall that the very reason we needed to become pluralists in the first place was the problem of statistical versus unidentified lives. That is, if our moral theory consisted of *nothing but* (*ex ante*) partially aggregative principles, then we would be required to save Gareth in **Mineshaft** even when doing so came at the expense of many thousands of unidentified miners. If we assign extreme weight to the claims reasons when compared with the goodness reasons, we arrive at roughly the same practical problem. We are required to save Gareth even when doing results in innumerable (unidentified) deaths.

My speculative suggestion here is that partial aggregationists thus face a kind of dilemma. On the one hand they can accept that partially aggregative principles belong to a pluralist theory, where this pluralist theory nevertheless proves to be exceptionally demanding, for it acknowledges weighty 'goodness facts' which compel demanding requirements. On the other hand they can accept that partially aggregative principles belong to a pluralist theory, but deny that, within such a theory, the 'goodness facts' compel demanding altruistic requirements. If that is right, then the partial aggregationist must exhibit an extreme bias in favour of identified lives in cases such as **Mineshaft**. Neither of these results, as things strike me, seem an especially appealing place to land.

Of course, the partial aggregationist could then try to resolve the overdemandingness concern by introducing *other* tools: by maintaining, for instance, that there is an absolutist limit to the quantity of sacrifice that givers can be required to make over the course of their lives as a whole, or by maintaining that personal prerogatives greatly enlarge the importance of your own moral projects within your own moral calculus.[56] In some sense, however, the game is then up. For the hope which I've so far been exploring was whether partially aggregative principles could furnish us with some kind of *independent traction* on the demandingness problem. In retreating to other tools, we would be giving up on this hope. Furthermore, the problems with available tools are well-known, such that it is hard to view there as being a consensus on which 'tool' is the one to use, or on whether any particular tool is fully intuitively adequate as a solution to the problem of overdemandingness.[57] Partially aggregative standards, it seems, would be of no help in helping us navigate through this difficult mire.

8. Conclusion

Let's recap. I've argued that partially aggregative principles systematically generate permissions for individuals to *not* donate towards charitable causes. This result has held for multiple species of partial aggregation. And it has held for multiple avenues of charity, including both statistical and identified ones.

At least one natural interpretation of this result is to say that partially aggregative principles thereby facilitate progress on the overdemandingness problem: such principles act as a moral shield, as it were, against the exacting demands of an impartial conception

[56] As in Scheffler (1982).
[57] For extensive discussion on concerns about the intuitive adequacy of absolutist aggregative limits, see Noggle (2009). For serious recent problems on prerogatives as a solution to demandingness, see Mogensen (2021). For important recent contributions on lifetime sacrifice as prerogative amplifier, see Pummer (2023).

of beneficence. For anyone who endorses this more optimistic interpretation, I've sketched two problems: a puzzle and a dilemma. The puzzle is that partially aggregative principles only furnish us with a highly modally fragile solution to demandingness. The dilemma is that, insofar as we embed partially aggregative principles within a pluralist theory, as it seems we should, we are left to choose between either a highly demanding conception of beneficence or a particularly vicious version of the identified lives bias. Making progress on these problems, I suspect, will demand progress on some underexplored questions about modal fragility and about the shape of the proposed pluralism.

References

Ashford, E. (2003) 'The Demandingness of Scanlon's Contractualism', *Ethics*, 113(2): 273–302.

Bostrom, N. (2013) 'Existential Risk as a Global Priority', *Global Policy*, 4: 15–31.

Bourguignon, L. & Mossé, M. (2025) 'How to Count Sore Throats', *Analysis*, 85(2): 1–11.

Budolfson, M. (2019) 'The Inefficacy Objection to Consequentialism and the Problem with the Expected Consequences Response', *Philosophical Studies*, 176(7): 1711–1724.

Chappell, R. Y. (2019) 'Willpower Satisficing', *Noûs*, 53(2): 251–265.

Cullity, G. (2003) 'Asking Too Much', *The Monist*, 86(3): 402–418

——— (2004) *The Moral Demands of Affluence*. Oxford University Press.

Curran, E. (2025) 'Longtermism and Aggregation', *Philosophy and Phenomenological Research*, 110(3): 1137–1151.

Dunn, E., Aknin, L. B., Norton, M. I. (2008) 'Spending Money on Others Promotes Happiness', *Science*, 319: 1687–1688.

Frick, J. (2015) 'Contractualism and Social Risk', *Philosophy & Public Affairs*, 43(3): 175–223.

Goodin, R. (2009) 'Demandingness as a Virtue', *The Journal of Ethics*, 13(1): 1–13.

Harbaugh, W., Mayr, U., & Burghart, D. R. (2007) 'Neural Responses to Taxation and Voluntary Giving Reveal Motives for Charitable Donations', *Science*, 316(5831): 1622–1625.

Hart, J. (2022) 'Tie-Breaks and Two Types of Relevance', *Ethical Theory and Moral Practice*, 25(2): 1–20.

——— (2025) 'Limited Aggregation's Non-Fatal Non-Dilemma', *Australasian Journal of Philosophy*, 103(2): 433–448.

Hawthorne, J., Isaacs, Y., & Littlejohn, C. (2023) 'Absolutism and Its Limits', *Journal of Moral Philosophy*, 21(1–2): 170–189.

Hills, A. (2003) 'Duties and Duties to the Self', *American Philosophical Quarterly*, 40(2): 131–142.

——— (2010) 'Utilitarianism, Contractualism, and Demandingness', *The Philosophical Quarterly*, 60(239): 225–242.

Hooker, B. (2000) *Ideal World, Real Code*. Oxford University Press.

Horton, J. (2017) 'Aggregation, Complaints, and Risk', *Philosophy & Public Affairs*, 45(1): 54–81.

——— (2018) 'Always Aggregate', *Philosophy & Public Affairs*, 46(2): 54–81.

——— (2020) 'Aggregation, Risk, and Reductio', *Ethics*, 130(4): 514–529.

Jackson, F. (1991). 'Decision-Theoretic Consequentialism and the Nearest and Dearest Objection', *Ethics*, 101(3): 461–482.

Johannsen, K. (2017) 'Animal Rights and the Problem of r-Strategists', *Ethical Theory and Moral Practice*, 20(2): 333–345.

Kagan, S. (1989) *The Limits of Morality*. Oxford University Press.

——— (2011) 'Do I Make a Difference?' *Philosophy & Public Affairs*, 39(2): 105–141.

Kamm, F. (1993) *Morality, Mortality: Volume I: Death and Whom to Save from it*. Oxford University Press.

——— (2007) *Intricate Ethics*. Oxford University Press.

Lazar, S. & Lee-Stronach, C. (2019) 'Axiological Absolutism and Risk', *Noûs*, 53(1): 97–113.

MacAskill, W. (2022) 'Are We Living at the Hinge of History?', in J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan (eds), *Ethics and Existence: The Legacy of Derek Parfit*. Oxford University Press.

MacAskill, W., Mogensen, A., & Ord, T. (2018) 'Giving Isn't Demanding', in P. Woodruff (ed), *The Ethics of Giving: Philosophers' Perspectives on Philanthropy*. Oxford University Press.

Mann, K. (2021) 'The Relevance View: Defended and Extended', *Utilitas*, 33(1): 101–110.

——— (2022) 'Relevance and Nonbinary Choices', *Ethics*, 132(2): 382–413.

Mann, K. & Barry, C. (2025) 'Aggregation and Permitting Reasons', *Journal of Moral Philosophy*, 1–20.

Mogensen, A. (2021) 'Moral Demands and the Far Future', *Philosophy and Phenomenological Research*, 103(3): 567–585.

——— (2024) 'The Hinge of History Hypothesis: Reply to MacAskill', *Analysis*, 84(1): 47–55.

Mulgan, T. (2001) *The Demands of Consequentialism*. Oxford University Press.

Murphy, L. (1993) 'The Demands of Beneficence', *Philosophy & Public Affairs*, 22(4): 267–292.

——— (2000) *Moral Demands in Nonideal Theory*. Oxford University Press.

Nebel, J. (2022) 'Totalism Without Repugnance', in J. McMahan, T. Campbell, J. Goodrich, & K. Ramakrishnan (eds), *Ethics and Existence: The Legacy of Derek Parfit*. Oxford University Press.

Nefsky, J. (2011) 'Consequentialism and the Problem of Collective Harm: A Reply to Kagan', *Philosophy & Public Affairs*, 39(4): 364–395.

Noggle, R. (2009) 'Give Till it Hurts? Beneficence, Imperfect Duties, and a Moderate Response to the Aid Question', *Journal of Social Philosophy*, 40(1): 1–16.

Norcross, A. (2004). 'Puppies, Pigs, and People', *Philosophical Perspectives*, 18: 229–245.

O'Brien, G. (2023) *Wild Animal Suffering and Human Responsibility: Essays on the Ethics of Beneficent Intervention in Nature*. Doctoral Thesis. University of Oxford. Available at: https://ora.ox.ac.uk/objects/uuid:6c87ff7f-1342-4d1e-80e6-c43b83fdf659.

——— (2025) 'The Case for Animal-Inclusive Longtermism', *Journal of Moral Philosophy*, 22(3–4): 336–359.

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

——— (2011) *On What Matters: Vol 1*. Oxford University Press.

Portmore, D. (1999) 'Does the Total Principle Have Any Repugnant Implications?', *Ratio*, 12(1): 80–98.

Pummer, T. (2023) *The Rules of Rescue: Cost, Distance, and Effective Altruism*. Oxford University Press.

Scanlon, T. (1998) *What We Owe to Each Other*. Harvard University Press.

Scheffler, S. (1982) *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions*. Oxford University Press.

——— (1992) *Human Morality*. Oxford University Press.

Sinclair, S. (2022) 'Explaining Rule of Rescue Obligations in Healthcare Allocation: Allowing the Patient to Tell the Right Kind of Story About Their Life', *Medicine, Health Care and Philosophy*, 25: 31–46.

Singer, P. (1980) 'Utilitarianism and Vegetarianism', *Philosophy & Public Affairs*, 9(4): 325–337.

Sobel, D. (2007) 'The Impotence of the Demandingness Objection', *Philosophers' Imprint*, 7(8): 1–17.

Suikkanen, J. (2019) 'Ex Post and Ex Ante Contractualism: A Synthesis', *Journal of Ethics*, 23(1): 77–98.

Tadros, V. (2019) 'Localized Restricted Aggregation', in D. Sobel, P. Vallentyne, & S. Wall (eds), *Oxford Studies In Political Philosophy*, 5: 171-204. Oxford University Press.

Thomson, J. A. (2022) 'Relief from Rescue', *Philosophical Studies*, 179(4): 1221–1239.
Thomas, T. (2018) 'Some Possibilities in Population Axiology', *Mind*, 127(507): 807–832.
Thorstad, D. (2023) 'High Risk, Low Reward: A Challenge to the Astronomical Value of Existential Risk Mitigation', *Philosophy & Public Affairs*, 51(4): 373–412.
Timmerman, T. (2015) 'Sometimes There is Nothing Wrong with Letting a Child Drown', *Analysis*, 75(2): 204–212.
Tomlin, P. (2017) 'On Limited Aggregation', *Philosophy & Public Affairs*, 45(3): 232–260.
Unger, P. (1996) *Living High and Letting Die: Our Illusion of Innocence*. Oxford University Press.
Voorhoeve, A. (2014) 'How Should We Aggregate Competing Claims?', *Ethics*, 125(1): 64–87.
——— (2020) 'Healthy Nails versus Long Lives: An Analysis of a Dutch Priority-Setting Proposal', in N. Eyal, S. A. Hurst, C. Murray, S. A. Schroeder, & D. Wikler (eds), *Measuring the Global Burden of Disease: Philosophical Dimensions*. Oxford University Press.
Walen, A. (2020) 'Risks and Weak Aggregation: Why Different Models of Risk Suit Different Types of Cases', *Ethics*, 131(1): 62–86.
Williamson, P. (forthcoming) 'Relevant Claims and Causal Ramifications', *Australasian Journal of Philosophy*.
York, K. (2024) 'Moral Demandingness and Modal Demandingness', *Journal of Ethics and Social Philosophy*, 28(2): 151–170.
Żuradzki, T. (2019) 'The Normative Significance of Identifiability', *Ethics and Information Technology*, 21(4): 295–305.

5.
Constraints, Consent, and Time

Abstract. I defend *ex ante* deontology, the view that we may only do those harms which are in the expected interests of all. My primary aim is to answer the recent concern that *ex ante* deontology cannot adequately handle choices made over time. In brief, I suggest that sophisticated versions of *ex ante* deontology deliver the right verdicts for the right reasons: they vindicate the key idea that we must never lethally and non-consensually impose ourselves upon the body of one for the sake of others. I conclude by considering whether we should prefer an *ex ante* deontological model or an *ex post* one. I suggest that, although recent arguments against *ex post* deontology have generally gone awry, we do indeed have good theory-neutral reasons for preferring the *ex ante* deontological model: the intuitions underpinning the *ex post* model are more easily explained away.

1. Introduction

1.1 Good Harms, Bad Harms

Some acts *do harm*. In virtue of doing harm, those acts qualify as impermissible. This can be the case even when those acts significantly promote the interests of other persons. By way of example, consider the familiar:

> **Footbridge.** Alice and Blaise are down on the tracks. An oncoming trolley will shortly kill them. You can save Alice and Blaise by pushing Charlie over the edge of a conveniently placed footbridge. Do so and his crushed, lifeless body will halt the trolley.[1]

[1] Foot (1967), Thomson (1985).

To push Charlie, in **Footbridge**, is to do him harm without his consent for the sake of benefitting persons who are not Charlie. In virtue of this fact, pushing Charlie seems to qualify as impermissible. To put things in the standard terms: pushing Charlie violates a *deontic constraint* against the doing of harm.[2]

A complication: some acts which do harm are nevertheless in the expected interests of every single person. Thus consider:

> **Suitcases**. As in **Footbridge**. But this time Alice, Blaise, and Charlie have been zipped shut in three separate suitcases and shuffled around. One suitcase is on the bridge and two others are down on the tracks. Nobody – including the patients themselves – knows which person is where. You can pull a lever which will instantly shunt the 'bridge suitcase' down onto the tracks, crushing that suitcase but shielding the two others.[3]

> **Suitcases (One Way Fabric)**. As in **Footbridge**. But Alice, Blaise, and Charlie have been zipped shut in three separate suitcases and shuffled around. One suitcase is on the bridge and two others are on the tracks. You don't know which person is where. But Alice, Blaise, and Charlie – *they* each know where they are, being able to see out of the one-way fabric from which their suitcases are made. You can pull a lever which will instantly shunt the 'bridge suitcase' down onto the tracks, crushing that suitcase but shielding the two others.[4]

Many feel greater uncertainty when confronted with either **Suitcases** or **Suitcases (One Way Fabric).** A common narrative goes as follows. These cases are harder to evaluate because they contain a *normative tension.* On the one hand the deontic constraint against doing harm speaks against pushing. (After all, if you push, then you will end up killing one person for the sake of two others.) On the other hand, a reasonable concern for the well-being of each individual person speaks in favour of pushing. (After all, if you push, then you will increase each individual person's odds of survival from 1/3 to 2/3, according

[2] Since Foot (1967), a familiar debate persists over whether it is doing or intending harm that matters; throughout, I bypass this debate by focusing upon harms which are both done and intended.
[3] Hare (2016). Three seminal discussions of cases with this broad structure were Harris (1975), Glover (1977: 100), and Kamm (1996: 291).
[4] This is the structure examined in Setiya (2020).

to your best available subjective probabilities.[5]) In light of this normative tension, what ought you do?

1.2. Some Common Intuitions

Let me start by simply setting out my own intuitions concerning the above three cases. My own intuitions are that:

(1) In **Footbridge**, it is impermissible to push Charlie.
(2) In **Suitcases**, it is required that you push whichever suitcase is on the bridge.
(3) In **Suitcases (One Way Fabric)** it is permissible to push whichever suitcase is on the bridge, but also permissible to refrain from doing so.

These intuitions certainly aren't universal. Nonetheless, they represent a reasonably common crop. They are, for instance, roughly the set of intuitions reported in recent work by Kieran Setiya.[6] If I were to try and explain why my intuitions change from case to case, I might say the following.

For familiar deontological reasons, it strikes me as impermissible to push Charlie in **Footbridge**. To push Charlie in **Footbridge** is to inflict harm upon Charlie without his consent for the sake of persons who *are not Charlie*. Things might have been different, perhaps, if we expected that Charlie wanted us to use his body as a human sacrifice in such a way. But *ex hypothesi*, this isn't the case. So, pushing Charlie appears to be impermissible.

In **Suitcases**, things seem different. In **Suitcases**, it is dramatically in everyone's expected interest that I push the suitcase. Furthermore, at the moment of choice, things look this way to all of us. This suggests that I may have (something like) presumed consent from each of Alice, Blaise, and Charlie to push the suitcase. Since I have

[5] Throughout, I assume that the best subjective probabilities are the ones which ought to be used in the calculation of prospects. For relevant discussion and complications, see Oberdiek (2017: Chapter 1).
[6] Setiya (2020).

(something like) presumed consent from each, this suggests that I ought to push the suitcase. What, though, about the familiar deontic constraint against doing harm? I find it much harder to believe that this deontic constraint could have much bite in **Suitcases** if there is indeed presumed consent from every single person in favour of performing the (expectedly beneficial) act in question.[7]

In **Suitcases (One Way Fabric)**, like Setiya, I am inclined to think that there is a genuine normative tension at play.[8] On the one hand, pushing is my chance to act out of reasonable welfarist concern*:* pushing is in the expected interests of every single person who stands to be affected by my decision, given all I can possibly know. On the other hand, to push is to inflict harm upon a particular person against their express wishes: if I push, then I will certainly kill someone (Charlie, say) who at this very moment refuses me permission to kill them in such a way, someone who would stop me pushing if they could.

As to how the conflict between concern and consent is to be resolved in **Suitcases (One Way Fabric)**, I'm not quite sure. What strikes me as plausible is that the tension between beneficence and autonomy might generate a permission leaning in either direction.[9] But perhaps it is most choiceworthy, all things considered, to refrain from pushing.

---

[7] For debates on actual versus hypothetical consent, see Rawls (1971), Dworkin (1972), Gauthier (1987), and Stark (2000). Of course, some may deny that the case contains hypothetical consent. For instance, ordinary persons sometimes *don't* consent to policies which are in their expected interests, as in Glover's (1977: 100) bomber pilots who allegedly rejected a suicide mission which was in the expected interests of all. My sense, as I discuss in §3, is that ordinary persons sometimes refrain from consenting to policies such as Glover's because those policies contain future points in time at which we will have to take identified persons and actively kill them against their future wishes – a feature absent in **Suitcases**. N.B. Kowalczyk (2022: 8) very briefly raises counterexamples to the consent strategy in which (say) children are within the suitcases, themselves lacking the autonomous capacity to consent, but it is hard to know what to make of these mournful counterexamples without saying a great deal more – a passing medical doctor may presumably save a sick child passed out on the street by interacting with their body, under the presumption of consent from the relevant parties (parents or guardians, say).

[8] Setiya (2020: 66-68).

[9] *Ibid.*

1.3 Ex Ante Deontology

Is there a general ethical model which accommodates the above intuitions? Fortunately, there is. According to *ex ante deontology*, whether or not deontological constraints apply depends upon the nature of the decision context. In those contexts where a given act φ is in the expected interest of every single person, one is required to φ. This is true even if φ demands that you perform a bad-kind killing. Outside of such contexts, however, the constraint against doing harm is operative as normal.

In order to capture my preferred verdict in **Suitcases (One Way Fabric)**, we may need to extend this nascent statement of *ex ante* deontology further. We may need to say something like the following: in those contexts where φ is in the expected interests of every single person (relative to my own subjective probabilities) but in which φ is known to harm someone against their own express wishes at the moment of choice, I have 'concern-based' moral reasons in favour of φ and 'consent-based' moral reasons against the performance of φ. It is then a substantive normative question about how the balance of these conflicting reasons is to be resolved.

Many, including myself, find *ex ante* deontology deeply appealing.[10] Unfortunately, it is well known that *ex ante* deontology faces distinctive problems when we turn our attention to choices made over time.[11] The problem has been pressed with especially great force in recent work by Kacper Kowalczyk.[12] My central aim, in what follows, is to defend *ex ante* deontology from these accusations: I aim to show that there is nothing wrong with being an *ex ante* deontologist over time.

[10] Those who draw a normative distinction between **Footbridge** and **Suitcases** include, for instance, Hare (2013; 2016) and Setiya (2020). Kowalczyk (2022: 4) suggests that *ex ante* deontology is the most desirable of the risk-sensitive deontological alternatives. *Ex ante* deontology also turns out to have independent usefulness, as in Unruh (2023).
[11] Influential cases were provided in Kamm (1996: 291–302).
[12] Kowalczyk (2022).

Of course, importantly, certain card-carrying deontologists prefer an *ex post* deontological model, a model according to which it is straightforwardly impermissible to push in all three of the above cases.[13] I conclude the discussion by responding to the *ex post* deontologist. My broad suggestion is that, although recent arguments against *ex post* deontology have generally gone awry, we do indeed have good theory neutral reasons for preferring the *ex ante* deontological model over than the *ex post* one. The intuitions which underpin the *ex post* model, I suggest, are more easily explained away.

2. The Diachronic Objection to Ex Ante Deontology

2.1 The Diachronic Objection to Ex Ante Deontology

The diachronic objection to *ex ante* deontology says that there is simply no plausible way of extending *ex ante* deontology such that it governs diachronic contexts – that is, such that it governs choices made over time. In light of this result, goes the objection, *ex ante* deontology must therefore be abandoned.

We can begin to grasp the diachronic objection by considering Frances Kamm's famous:

> **Poison Soup.** Three of us have just eaten a lovely supper. Oh dear! Two of the meals were poisoned and we don't know which is which. Never fear. The third meal contained an antidote – indeed, two doses of antidote. Right now, we can all agree to the following arrangement (the 'murder arrangement'). In an hour's time, when two of us begin to look peaky, those two may cut open the stomach of the third in order to access the precious antidote contained within.[14]

Cases like **Poison Soup** are distinctive insofar as they contain a time delay. We are currently clueless about who has swallowed the antidote. But this won't always be the

[13] As in Kamm (1996).
[14] Kamm (1996: 291).

case. In an hour's time, we will each know who is on course to live and who is on course to die.

Cases like **Poison Soup** pose something of a puzzle for *ex ante* deontologists. After all, on the one hand, *ex ante* deontology seems to require that you commit to the murder arrangement now: the murder arrangement is currently in everyone's expected interests. On the other hand, *ex ante* deontology seems to require that you back out of the murder arrangement in an hour's time: in an hour's time, the murder arrangement *won't* be in everyone's expected interests. It isn't puzzling *per se* that your moral obligations can change over time. What *is* quite puzzling, though, is that *ex ante* deontology seems to require that you make an agreement which you know you must immediately recant.

There are at least three natural ways in which we might extend *ex ante* deontology to cover diachronic cases like **Poison Soup**. We could endorse *myopic ex ante deontology*, *resolute ex ante deontology*, or *sophisticated ex ante deontology*.[15] The proponent of the diachronic objection to *ex ante* deontology maintains, of course, that none of these three separate models turns out to hold plausibility. Let me set out the models one by one.

According to the simplest model, *myopic ex ante deontology*, we should simply allow ourselves to be carried along by the passage of time, accepting (or refusing) a harmful policy on the basis of whether or not it is currently (currently not) in the expected interests of all. The murder arrangement in **Poison Soup** is initially in everyone's expected interest, so initially, we must agree to it. But the murder arrangement will not be in everyone's expected interest in an hour's time, so in an hour's time, we must back out of it.

[15] This trichotomy is ubiquitous: for an accessible statement and experimental work on the degree to which it tracks real-world approaches to choice over time, see Hey & Lotito (2009).

It's not hard to see why the myopic model is unappealing. The myopic model requires that we commit to the murder arrangement, wait an hour, and then back out of it. But not only is this shortsightedness. It is shortsightedness which can easily lead to gratuitous badness.[16] Suppose, for instance, that there is some small cost involved in agreeing to the murder arrangement now but backing out later – doing so means that the single surviving 'antidote swallower' ends up melancholy, having to spend an additional hour thinking about the unpleasantness of murder. If this is right, then the myopic model leads to gratuitous badness. It requires that we leave the single antidote swallower alive and melancholy rather than alive and happy. This is an extremely difficult consequence to accept. (The proper language: the myopic model turns out to violate the *Sequential Pareto* principle, by requiring that we perform a sequence of actions resulting in an outcome which is guaranteed worse for some and guaranteed better for none.)

This unfortunate result is represented neatly in the following diagram.[17] The squares represent your potential choice points within **Poison Soup**, and the circles represent the moment at which your uncertainty is resolved – that is, the moment at which you learn who has swallowed the antidote. The myopic model requires that we go up one branch by agreeing to the murder arrangement now. It then requires that, in an hour's time, however our uncertainty ends up being resolved, we back out of the murder arrangement. This leaves the single survivor melancholy (red). By contrast, an alternative was to simply reject the murder arrangement from the outset, leaving the single survivor happy (green):

[16] Kowalczyk (2022: 6) makes these all these points, employing a novel variant of **Suitcases** to do so.
[17] *Ibid.*

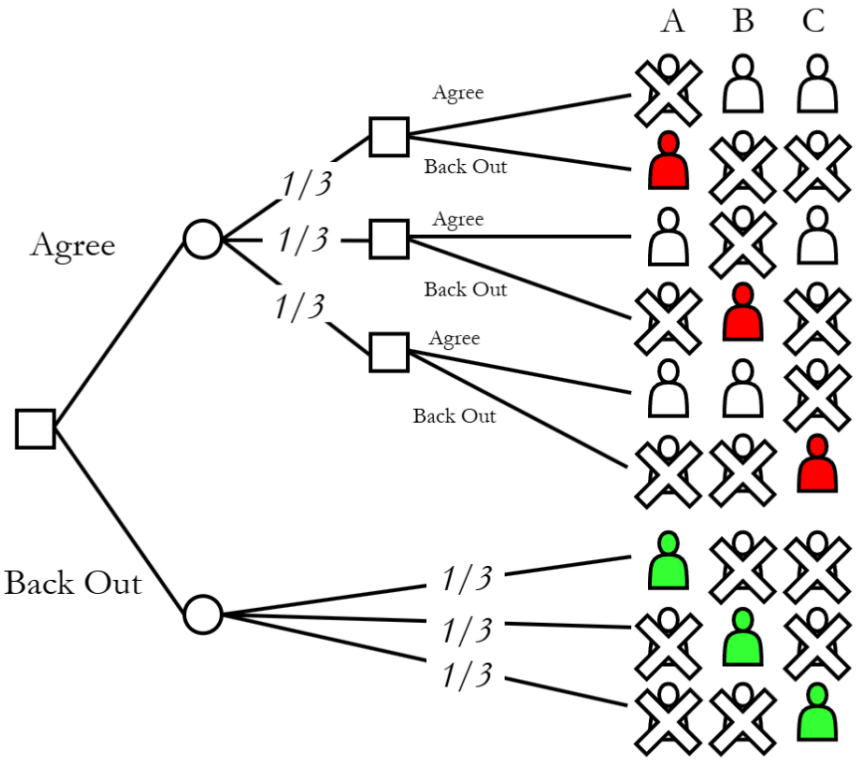


Fig 1. ***Poison Soup*** *(with backing out penalty)*

Already, then, it seems we must turn our attention to some of the other models on offer. According to *resolute ex ante deontology*, we should approach **Poison Soup** by binding ourselves to the mast.[18] Here is how the reasoning works. Currently, it is maximally in everyone's expected interest to agree to the murder policy and to carry through with it (to go up an 'agree' branch at each stage). What's more, it is currently *against* everyone's expected interest to make the murder agreement, wait an hour, and then back out (to choose the first 'agree' branch and to subsequently choose a 'back out' branch). This suggests that we must do the following: we must commit ourselves to the murder arrangement and make ourselves resolute to it. When we later find ourselves with a chance to back out from the murder arrangement, we must not do so.

[18] For canonical work on the general distinction between resoluteness and sophistication, see Strotz (1955), Machina (1989) and McClennan (1990).

Unfortunately, I know of exceedingly few deontologists who are intuitively happy with the verdicts generated by resolute *ex ante* deontology.[19] Why? Well, in general, resolute *ex ante* deontology requires that we commit ourselves to doing bad-kind killings 'down the line' even though those future killings are not in the expected interests of victims at the time at which they are carried out. For anyone who holds even remotely deontological sympathies, this is a very hard pill to swallow.[20]

Suppose for instance that we regulate our conduct in accordance with resolute *ex ante* deontology in **Poison Soup**. This means the following. We all agree to the murder arrangement at t_1. The clock rolls around. At t_2, we learn which one of us swallowed the antidote – Freddy, say. Freddy now pleads with us desperately to spare his life. Nevertheless, the two of us are unmoved. We solemnly remind Freddy that this was a plan to which he previously agreed. We turn towards Freddy, raise our knives, prepare to land the felling blow … Very few deontologists are willing to accept, I've found, that the two of us do something morally permissible (let alone *morally obligatory*) in killing Freddy.

If myopic and resolute models are to be rejected, then we've already arrived at the remaining model constituted by *sophisticated ex ante deontology*.[21] According to sophisticated *ex ante* deontology, we should approach cases like **Poison Soup** by engaging in backwards induction: that is, we should skip forward to future choice points and ask how we *would* be constrained to act in those choice points, conditional on conformity with deontological constraints. We then ask how we should act now, in light of our future anticipated behaviour.

---

[19] As Kowalczyk (2022: 15) notes, something like the resolute view seems to have been temporarily endorsed by Thomson (1990: 176–202).

[20] There are other general problems facing resolute models, such as the problem of figuring out when plans commence, and by extension, which are the optimal plans to which we should make ourselves resolute. See, for instance, discussion in Gustafsson (2022: 66–74). I find none of these problems quite as worrying as the fact that resolute *ex ante* deontology yields counterintuitive verdicts in cases like **Poison Soup**.

[21] Again, on the general distinction, see Strotz (1955), Machina (1989) and McClennan (1990).

Think of things the following way. Let's suppose, in **Poison Soup**, that we *did* agree to the murder policy right now (going up the initial 'agree' branch). In an hour's time, after the revelation of identities, the murder arrangement would then involve doing harm to someone against their expected interests. We would thus find ourselves forced to back out of the murder arrangement (that is, to go down a future 'back out' branch). So, we should really evaluate our choice in the following way. Our first option is to agree to the murder arrangement now *and* back out later (leaving the single antidote-swallower melancholy / red). Our second option is to simply never agree to the murder arrangement in the first place (leaving the single antidote-swallower happy / green). Since this is our choice, we should simply never commit to the murder arrangement in the first place. We should leave the antidote swallower – whoever they end up being – alive and happy rather than alive and sad.

We might think that the prospects for sophisticated *ex ante* deontology look rather good. In particular, sophisticated *ex ante* deontology gets exactly the right answer in **Poison Soup**: we must simply eschew the murder arrangement from the outset. In delivering this verdict, sophisticated *ex ante* deontology avoids the twin evils of shortsightedness (committing to the murder arrangement now and backing out later) and of resoluteness (binding ourselves to commit murder at some future time).

Kowalczyk has recently argued, however, that the sophisticated model nonetheless faces serious challenges. They illustrate with the following variation of **Suitcases**:

> **Suitcases II.** As in **Suitcases**, except for the following. It's now 12:00 and you can currently press the 'push' lever. At 1:00 everyone's location will be revealed and you will have yet another chance to press the 'push' lever, if you haven't already done so. At 1:15, the trolley will pass under the footbridge. Unfortunately, there is a small cost involved in pushing the lever early. Doing so will leave the two track-folk worse-off by some small margin: they will become melancholy (red rather than green), having to listen to the unhappy cries of whoever has been sacrificed for a whole additional hour.[22]

[22] Kowalczyk (2022: 10).

Sophisticated *ex ante* deontology seems to struggle in **Suitcases II**. Why? Well, for reasons to be explained momentarily, sophisticated *ex ante* deontology requires that you push at 12:00. But this is to violate Sequential Pareto just as myopic *ex ante* deontology did: it is to require that we perform a sequence of acts which leaves some worse-off and none better-off. After all, pushing at 12:00 brings about the exact same outcome as does pushing at 1:00, except for the single fact that the two survivors (whoever they are) end up melancholy rather than happy:

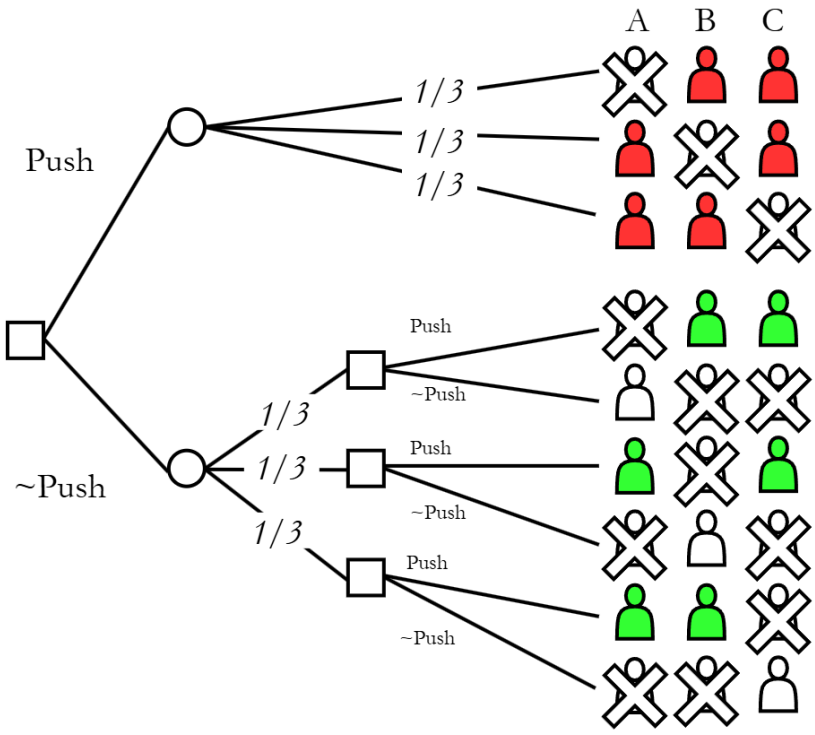


Fig 2. ***Suitcases II*** *(with hastiness penalty)*

Why, in brief, does sophisticated *ex ante* deontology require that you push at 12:00 rather than push at 1:00? Well, again, the sophisticated model operates via the use of backwards induction. Suppose that you *didn't* push now (going down the initial ~*Push* branch). Then, at 1:00, you would also have to refrain from pushing (go down the future ~*Push* branch).

After all, pushing at 1:00 would be against the expected interests of whoever is later revealed to be upon the bridge. So your first option in **Suitcases II** *really* consists of performing the sequence of acts 'don't push at 12:00' and 'don't push at 1:00.' This is certain to leave the single person on the footbridge alive. Your second option is to simply push now, leaving two persons alive but slightly melancholy. Which of these options is better? From the current vantage point, it is strongly in everyone's expected interests that you *Push* now, given that this leaves two (as yet unidentified) persons alive and slightly melancholy rather than one (as yet unidentified) person alive upon the footbridge. So sophisticated *ex ante* deontology orders you to push now. This is even though pushing now is a violation of the Sequential Pareto principle, leaving some worse-off and none better-off.

This final result would seem to complete the diachronic objection against *ex ante* deontology. Since the sophisticated model *also* has unacceptable implications, just as the myopic and resolute models did, *ex ante* deontology must be abandoned altogether. There is no plausible way of extending *ex ante* deontology such that it can govern choices made over time.

2.2 In Favour of Sophisticated Ex Ante Deontology

My general response to the diachronic objection goes as follows: is it in fact troubling that sophisticated *ex ante* deontology violates Sequential Pareto in **Suitcases II**? Kowalczyk certainly takes this result to be deeply problematic, or at the very least, to serve as part of a cumulative case against sophisticated *ex ante* deontology. I am not quite convinced, however.[23]

[23] Kowalczyk (2022: 12) also objects, for instance, that the sophisticated model leads to implausible information avoidance. In general, the concern is that the sophisticated model forces you to "actively resist" becoming better informed in cases like **Suitcases II**, even when doing so leaves someone unnecessarily

It's certainly true that, if you push at 12:00 in **Suitcases II**, you do something undesirable. You violate Sequential Pareto by bringing about a state of affairs which is worse for some and better for none. Nonetheless, there's also a transparent sense in which you do something *highly desirable*. You select a course of action which, at no point in time t_i, requires violating someone's bodily autonomy against their express wishes at t_i. By the lights of any deontologist, this is an extremely desirable feature for a sequence of acts to have.

Things are the other way around, of course, if you wait to push at 1:00. If you wait to push at 1:00, then you satisfy Sequential Pareto. Nonetheless, you do something lamentable. You violate someone's bodily autonomy at some moment t_i, against their express wishes at t_i. To be precise, you push (say) Charlie once it is revealed that he is the person on the footbridge, even though (say) Charlie is currently asking you to spare his life and refrain from making his body into a human train-stop.

Now, in independent cases, when the Pareto optimal outcome can only be secured by violating someone's bodily autonomy without their consent at some time slice t_i, deontologists are regularly inclined to uphold considerations of bodily autonomy rather than to uphold Pareto principles. Take the ubiquitous example in which a dying patient is in urgent need of a blood transfusion but in which they refuse the transfusion on religious grounds (**Transfusion**). We can imagine that, on balance, providing the blood transfusion would be overall better for the patient *and* just as good for their family. Nonetheless, deontologists systematically judge that you must refrain from delivering the blood transfusion: you must, in the end, forego the Pareto optimal outcome. Furthermore,

worse-off. I do not think this can be a probative objection against *ex ante* deontology, simply insofar it turns out to be unbelievably difficult to avoid information avoidance across a wide range of plausible views concerning rationality and ethics. As Maher (1992) shows, for instance, there are circumstances in which both evidential and causal decision theories judge it irrational for utility maximisers to make cost-free observations.

this judgement turns out to be robust under a range of salient variations. The judgement persists even if the religious belief, say, was only formed a year ago. It persists even if the transfusion can be delivered surreptitiously such that the patient never learns that they have received it. And it persists even if the patient's desire isn't grounded in religious beliefs or values *per se*, but is simply a desire with which the patient reflectively identifies.[24]

If this is right, however, then **Suitcases II** cannot constitute a substantive objection to sophisticated *ex ante* deontology. Deontologists were *already happy*, even in synchronic cases, to give up the Pareto optimal outcome if such an outcome could only be secured by violating someone's bodily autonomy at some time t_i against that person's express wishes at t_i. If deontologists were already happy to give up the Pareto optimal outcome in the synchronic case (as a means of respecting bodily autonomy), then it can't be viewed as a substantive further objection to *ex ante* deontology that it yields this implication in the diachronic case (as a means of respecting bodily autonomy). That, in short, is the nub of the reply I want to make to the objection posed by **Suitcases II**.

Of course, already, some important complications have emerged. Let me address one of these now so that I can set it aside. One live possibility is that something like 'bodily autonomy' is a constituent component of welfare, such that having one's bodily autonomy violated against one's express wishes, at some time t_i, makes one's life go especially poorly.[25] Perhaps, for this reason, providing the blood transfusion in **Transfusion** *would indeed* make the patient's life go all-things-considered worse. In this way we would still conform to Pareto principles by refraining from giving the patient a blood transfusion: we

[24] Perhaps as in Frankfurt (1971).
[25] This concern is footnoted by Kowalczyk (2022: 4). For one recent statement of autonomy as a welfare good, see Hooker (2015).

would do what is best *for the patient*, even in denying them many happy years of future life.[26]

This is certainly one possible explanation of why it would be wrong to give the patient a blood transfusion in **Transfusion**. Nonetheless, if this style of explanation is right, then sophisticated *ex ante* deontology can no longer be accused of violating Sequential Pareto in **Suitcases II** after all.

How so? Well, in **Suitcases II**, sophisticated *ex ante* deontology requires that I push at 12:00 rather than pushing at 1:00. At first glance this might *look* Pareto inferior when contrasted with pushing at 1:00. After all, pushing at 12:00 is worse for the two surviving track folk, and the unfortunate 'bridge person' is going to die in either case. Nonetheless, under the current proposal, pushing at 12:00 *would indeed* be better for a particular person. It would in fact be better for the 'bridge person' who ends up dying. For if you push at 12:00, then you kill 'bridge person' with their blessing, since neither you nor 'bridge person' know where they are. By contrast, if you wait to push at 1:00, after locations have been revealed, then you end up killing 'bridge person' (Charlie, say) in a way which violates their bodily autonomy against their express wishes at the moment of choice, pushing Charlie when Charlie himself does not wish to be pushed. Under the current proposal, this kind of bodily autonomy violation *just is* the sort of thing which makes someone's life go especially poorly with respect to overall welfare. Since there is someone whose life ends up going *better* if you push at 12:00, sophisticated *ex ante* deontology can no longer be accused of violating the Sequential Pareto principle in **Suitcases II** – under, at least, the above proposal that bodily autonomy violations are best understood as hits to welfare.

[26] As mentioned, perhaps these future years could contain blissful ignorance via a surreptitious transfusion delivery.

2.3 The Reply from Irrevocable Consent

I've suggested that the prospects for sophisticated *ex ante* deontology look surprisingly good. Sophisticated *ex ante* deontology might indeed require that we violate Sequential Pareto in a case like **Suitcases II**. But this would only hold insofar as bringing about the Pareto superior outcome, in a case like **Suitcases II**, required that we violate someone's bodily autonomy (at some time t_i) against their express wishes at (at t_i). Deontologists *already* believed, even in the synchronic case, that we must forego Pareto optimal outcomes which can only be brought about by enacting bodily autonomy violations of such a kind. Thus, the current objection cannot serve as a substantive mark against sophisticated *ex ante* deontology.

I now want to consider a potential reply to my above defence. In general terms, my above defence drew upon what we might call the:

> *Bodies Prohibit Best Principle.* We may only perform a body-violating act which brings about a Pareto superior outcome insofar as we have *normatively forceful consent* from the person whose body is being violated by that act, at the time of performance.[27]

The Bodies Prohibit Best Principle explains, at least in part, why it is wrong to deliver the blood transfusion in **Transfusion**. Perhaps providing a blood transfusion would indeed be better for the patient and better for others. Nonetheless, we do not currently have normatively forceful consent from the patient to enact the transfusion. Thus, we may not enact the transfusion. The phrase 'normatively forceful consent' is a term of art. But the general idea here is familiar. There are many kinds of consent which we take to lack normative force – most famously, consent which is coerced, extracted, or brainwashed.

[27] The principle is most plausible when glossed in moderatist (rather than absolutist) terms. Clearly, if the bodily autonomy violation is very small (pinching an arm) and the pay-off is very great (saving the life of a third party) we ought to enact the violation. Nonetheless, I take something like the *Bodies Prohibit Best Principle* to explain why the sophisticated judgement is warranted in examples such **Suitcases II**.

Forms of consent such as these would not generate permissions, I take it, in favour of providing the blood transfusion in **Transfusion**.

The reply in which I'm currently interested goes as follows. Suppose, in **Suitcases II**, that I do indeed push the sacrificial suitcase at 1:00 after the revelation of everyone's respective location. You might nonetheless think that I *do* have normatively forceful consent from whomsoever I push at 1:00, that I push them in such a way. If this is right, then sophisticated *ex ante* deontology would still do something highly defective in requiring that I push at 12:00. Sophisticated *ex ante* deontology would violate Sequential Pareto, and we could not justify this violation by maintaining that pushing at 1:00 would have violated the Bodies Prohibit Best Principle.

How could this be the case? How could I have normatively forceful consent from (say) Charlie to push them from the footbridge at 1:00, even once we have all learned that Charlie is the one upon the footbridge? Well, of course, at 1:00, once we have learned everyone's respective location, Charlie doesn't want me to use his body as a train-stop. He is probably yelling at me not to do so. Nonetheless, at 12:00, the plan of "wait and push" was maximally in Charlie's expected interests. So perhaps, at 12:00, Charlie could have been expected to give *irrevocable consent* to the plan of action which involved pushing the suitcase at 1:00. If Charlie would have given irrevocable consent to this plan at 12:00, then perhaps I do indeed have normatively forceful consent from Charlie to push his body at 1:00, even once it is revealed that Charlie is in fact upon the footbridge.

This is a very natural reply to my above defence. Nonetheless, upon reflection, I remain unconvinced. The first point to make, here, is that it is a substantive and difficult question about whether there can even exist such a thing as irrevocable consent. Our intuitions about the legitimacy of irrevocable consent depend heavily upon the context.

In some contexts, such as marketplaces, it is very tempting to think that irrevocable consent is possible (and indeed, that irrevocable consent is partially constitutive of the context in the first place). Suppose that you and I have entered the following contract: I will be the exclusive wheat supplier for your bread factory. Nonetheless, in the extremely unlikely event that the weather downgrades the quality of my grain, I consent to pay you a premium to cover your lost revenue. Surely my consent to this contract must be understood as irrevocable. In the unlikely event that my crop *is* downgraded, I cannot turn around and say: "Fortunately, consent is always revokable. Now that conditions no longer favour me, I no longer consent to our arrangement. I will not pay your premium."

In such contexts there is a natural explanation of why contractual agreement must be irrevocable. It seems extremely instrumentally useful, for economic agents wishing to maximise their long-run accumulation of capital, to have conditions of exchange which are stable and predictable, such that each agent can reliably trust upon and plan with respect to the assurances given by others. These conditions are preserved precisely insofar as there is a legally enforced norm that contractual consent must be binding – that I may not revoke my consent simply because I rolled the dice and they happened to come up snake eyes.[28]

In other contexts, such as sexual agreements, it is very tempting to think that irrevocable consent simply *is not* possible.[29] Suppose that *A* solemnly promises sex to *B*. Furthermore, *A* explicitly stipulates that they are providing irrevocable consent to sex: *A* explicitly tells *B* that all of *A*'s future remarks on the topic are to be ignored. Nevertheless, when the moment comes, *A* sincerely changes their mind and expresses that they do not wish to have sex with *B*. *B*, recalling *A's* previous remarks, insists on having sex anyway.

[28] For recent defence of instrumentalism concerning contract law, see Murphy (2014).
[29] For recent discussion of sexual promises, see Liberto (2017).

This is not a consensual interaction; it is rape. It does not seem possible for past-*A* to give such a thing as irrevocable consent to *B* in regards to the use of future-*A*'s body for sex. This is true even if it was previously in the expected interests of *A* to voice an irrevocable promise such as this one.

Intuitively, cases like **Suitcases**, **Suitcases II**, and **Poison Soup** more closely resemble the latter context than the former. These are all cases in which we contemplate violating the physical bodily integrity of beneficiaries at some future point in time, even though we know that, later on, a given beneficiary will expressly reject our making use of their body in such a way. So it should not be surprising, then, if these are cases in which irrevocable consent simply turns out to be impossible. Particular evidence for this was found in our intuitive reaction to **Poison Soup**. The initial intuitive reaction to **Poison Soup**, I take it, was that it would be morally monstrous for us to wait until the revelation of identities and then gruesomely murder whoever it was who in fact swallowed the antidote (Freddy, say). This reaction would be inexplicable if these were in fact cases in which irrevocable consent was possible at t_1. For if that were the case, then we would indeed have *Freddy's* consent to murder him at t_2: we would have the irrevocable consent which Freddy gave to such a course of action at t_1.[30]

There are different possible explanations of why we do not seem to have normative powers in these cases, at t_1, to irrevocably consent to the use of our bodies at certain ways at some future time t_2. The question has been explored at length in recent work by Angela

[30] Even if irrevocable consent turned out to be possible in these cases, I am not sure this would get the diachronic objection to *ex ante* deontology across the line. For then what seems to me the most worrying concern with the resolute model – that it delivers intuitively unappealing verdicts – can be explained away. We *think*, intuitively, that the resolute model is unacceptable because it requires us to murder Freddy at t_2. But there can in fact be nothing wrong with this course of action, as this was something to which Freddy irrevocably consented at t_1. I would find this result somewhat unsettling, but not as unsettling as abandoning deontology altogether in cases like **Footbridge**.

Sun, amongst others.[31] For Sun, we cannot irrevocably consent in such a way because to do so would be to deny our future selves the capacity for *bodily integrity*, where bodily integrity consists, roughly, of an agent being able to make sense of (and to some degree control) the way in which her body *qua* vessel of agency relates to the external world. Our ordinary sense is that this is a fundamental capacity from which we do not have a moral permission to alienate our future selves. As Sun notes, however, this is not the only potential explanation of why consent must be revokable in bodily contexts. Another explanation, at least in certain cases, is that consent at t_1 cannot qualify as irrevocable because it tries to bind you at t_2 even though you currently lack knowledge about the character of your 'mental life' at t_2. Perhaps at 12:00 in **Suitcases II**, for instance, you cannot know *what it would feel like* to be facing with certainty a lethal bodily violation against your express wishes. Perhaps the fact that you lack this information somehow entails that you only have normative powers to revocably consent to the plan in which the decision maker will push at 1:00, rather than to irrevocably consent to such a plan.[32]

For current purposes, it's not obvious whether we need to commit to any particular one of these underlying explanations. This is because I already take cases like **Poison Soup** to provide fairly strong intuitive evidence that prior consent *cannot* be revocable in such cases, at least under our ordinary conception of consent. What matters for current purposes is that, given that irrevocable consent *doesn't* seem possible in these cases, my previous defence of sophisticated *ex ante* deontology retains a high degree of plausibility. Sophisticated *ex ante* deontology delivers the entirely acceptable verdict that you ought to push at 12:00 in **Suitcases II**. Doing so violates Sequential Pareto. But it is a means of

[31] Sun (2025). Relevant previous discussion is in Dougherty (2014).
[32] Sun (2025: 7–9). Sun makes the point with respect to independent cases – attempting to give irrevocable consent to bungee jumping, for instance – but I take it that the central reasoning would track, to at least some degree, to the cases considered here.

respecting the Bodies Prohibit Best Principle, which you would have violated had you pushed at 1:00.

3. *Ex Post* or *Ex Ante* Deontology?

I've argued that the diachronic objection against *ex ante* deontology is unsuccessful. In particular, sophisticated *ex ante* deontology delivers the right answers for the right reasons. If this is correct, there is nothing wrong with being an *ex ante* deontologist over time.

What I've so far skipped over, however, is that certain deontologists are not inclined to share my judgements across the three cases with which we began – across **Footbridge, Suitcases,** and **Suitcases (One Way Fabric)**. Certain *ex post deontologists* judge that it is straightforwardly impermissible to push in all three cases.[33]

My final task, then, is to argue that we should indeed prefer the *ex ante* deontological model over the *ex post* one. My general suggestion goes as follows. Although recent arguments against *ex post* deontology have generally gone awry, in the sense that they tend to lack dialectical force, there is indeed a promising and potentially theory-neutral way of demonstrating the superiority of the *ex ante* model. The trick is to show that the intuitions underpinning the *ex post* model are more easily explained away.

3.1. The Composition Argument

The most influential recent argument against hardline *ex post* deontology is provided by Caspar Hare.[34] To see how Hare's *Composition Argument* works, consider the following variant upon **Suitcases:**

[33] The most notable proponent being Kamm (1996).
[34] Hare (2016).

> **Suitcases (Parallel Tracks).** A runaway trolley is heading towards two suitcases. A third suitcase (a potential train stopper) is perched upon a conveniently placed footbridge. Alice, Blaise, and Charlie are within the suitcases, though nobody knows who is where. The exact same situation is repeated in parallel on a nearby track. However, the suitcases on this second track contain mere sand.
>
> You can pull any number of the following three levers – the *Move Alice* lever, the *Move Blaise* lever, and the *Move Charlie* lever. By pulling the *Move Alice* lever, you'll swap Alice with her corresponding 'sand suitcase' upon the second track (though you won't see the move). By pulling the *Move Blaise* lever, you'll swap Blaise with his corresponding 'sand suitcase' upon the second track (though you won't see the move). By pulling the *Move Charlie* lever, you'll swap Charlie with their corresponding 'sand suitcase' upon the second track (though you won't see the move). Pull any single lever, however, and the suitcase which is currently atop the second track's footbridge will be shunted down to serve as a train stop.[35]

The case is intricate and deserves a word of explanation. In **Suitcases (Parallel Tracks)**, you cannot shunt the suitcase off the first footbridge onto the first track. Nonetheless, you can 'swap out' each individual person from their position on the first track to a corresponding position on the second track. The moment that *anyone* is swapped out in such a way, however, the suitcase on that second track's footbridge is shunted down to serve as a train stop.

In even plainer English, when I deliberate about whether to pull the *Move Alice* lever, I might think to myself as follows: "Well, I don't know which suitcase Alice is in and I can't find out. But if I press *Move Alice* I'll certainly swap her suitcase, wherever it is, to a parallel position upon the second track. And I'll certainly shunt down the 'footbridge suitcase' on the second track, if it hasn't already been shunted. Of course, I'm not omniscient. But as things seem to me, pulling the lever is a pretty good deal *for Alice*. For it increases her probability of survival from 1/3 to 2/3. Furthermore, it does so without interfering with the body of anyone else."

[35] Hare (2016: 461–462).

Hare reasons about **Suitcases (Parallel Tracks)** in the following way. He thinks that you ought to pull the *Move Alice* lever independently of whatever else you ought to do. He thinks the same is true of each other *Move* lever. But if this is right, it seems you ought to pull all three *Move* levers – indeed, you ought to perform a composite act which would involve simultaneously pulling *all three Move levers*. You ought, that is, to do a bad-kind killing in **Suitcases (Parallel Tracks)**, killing whoever is currently on the footbridge for the sake of two others. Distilling these sentiments a little more formally, we arrive at:

> *The Composition Argument*[36]
> 1. In **Suitcases (Parallel Tracks)**, irrespective of what you do later, you ought to pull the *Move Alice* lever. Irrespective of what you did earlier, you ought to pull the *Move Blaise* lever, . . . , and so on.
> 2. For any composite action [A_1 . . . A_n], IF irrespective of whether you do [A_2 … A_n], you ought to do A_1, and . . . and irrespective of whether you do [A_1 . . . A_{n-1}], you ought to do A_n, THEN you ought to do [A_1 . . .A_n].
> C. In **Suitcases (Parallel Tracks)**, you ought to pull all three *Move* levers; you should move one person into the tracks such that their body is used as a train stop to save two others.

This argument would be a proof against *ex post* deontology because it should not make a difference to our evaluations, presumably, whether we are dealing with a case containing a single composite act (**Suitcases**) or a perfectly analogous case containing a set of decomposed sub-acts (**Suitcases (Parallel Tracks)**).

For reasons which have been briefly explored elsewhere, I am not persuaded by the Composition Argument.[37] In short, I think that premise (1) is highly questionable. It's not obvious to me whether you should press each of the *Move* levers independently of whatever else you do.

---

[36] *Ibid.*
[37] Very similar concerns are raised in Kowalczyk (2022: 18).

We can very powerfully explain why by considering some familiar and structurally analogous cases from the literature on collective action. Thus consider:

> **Lollipop of Doom.** Each of a million agents P_1 – $P_{1,000,000}$ can press their own respective button. Pressing the button will give that agent's child a lollipop. Pressing the button will also release a harmless chemical – from the set $CHEM_1$, $CHEM_2$ … $CHEM_{1,000,000}$, respectively – into the water supply of a foreign country. If all million chemicals are released, however, then there will be an interaction effect, and a new gaseous compound will rise into the air of that country, killing every single one of its million inhabitants.[38]

Imagine for a moment – however implausible their reasoning sounds – that the agent P_1 deliberated about their choice in **Lollipop of Doom** in the following way: "I ought to press my button. Furthermore, I ought to do so independently of whether anyone else presses their button. For the choice of whether or not to press my button is a private one: it affects none other than myself and my child."

We could very quickly point out to P_1 that they are making a horrendous mistake in moral reasoning. They neglect the fact that, if other persons press their buttons too, then P_1 will be the 'trigger' of an interaction effect such that a morally undesirable event ends up obtaining in the world at large. Instead, P_1 should evaluate the choiceworthiness of button-pressing as follows. They should specify the potential moral value which would arise from button-pressing under each possible state of affairs which might follow from their act. They should weight these moral values by the relevant probabilities. And they should sum the resulting values in order to arrive at an overall appraisal of the expected agent-relative choiceworthiness of button-pressing.

Suppose, for instance, that P_1, has narrowed things down to three remaining possibilities. Either all of the other persons will press their respective buttons (0.3 likely),

[38] Using Kagan's (2011) language, the example is a 'pure triggering case.'

half of the other persons will press their respective buttons (0.5 likely), or none of the other persons will press their respective buttons (0.2 likely). P_1 might then represent their choice in the following way, with numbers in each cell representing very rough (cardinally meaningful) estimates of the moral value or disvalue obtaining in each outcome:

Table 1. P_1's choice

	All others press (0.3)	Half of others press (0.5)	No others press (0.2)
Press button	-100,000,000,000	1	1
~Press button	0	0	0

The obvious thing to say is that, so long as there is a non-trivial chance that P_1*'s* decision to press the button will serve as the *triggering press* for the calamity – as represented by the first cell – then button pressing is going to qualify as the less choiceworthy option. P_1 ought not press.

If we endorse premise (1) of the *Composition Argument*, we are vulnerable to making the exact same species of mistake which P_1 made. For instance, suppose I pull the *Move Alice* lever. Suppose I justify my conduct by saying: "I ought to pull the *Move Alice* lever. Furthermore, I ought to do so independently of whether I pull any future levers. For the choice is a private one which affects none other than myself and Alice."

This is to make the exact same species of mistake, the deontologist can reply, that P_1 made in **Lollipop of Doom**. It is to ignore the possibility that there will be a morally undesirable interaction effect if I pull the *Move Alice* lever *and* later go on to pull other *Move* levers too, an interaction effect which involves other persons too. The *ex post* deontologist does indeed believe that such a morally undesirable interaction effect would emerge: they think there is something especially morally awful in taking one person's body and using it as a human train-stop to save the life of another.

As before, the correct way forward instead seems to involve assigning numbers to things. Suppose that, in **Suitcases (Parallel Tracks),** you've confidently narrowed things down to two broad possibilities. You know that either *all* of your future time slices will pull their respective *Move* levers (0.4 likely) or *none* of your future time slices will pull their respective *Move* levers (0.6 likely). (You know that you won't press just one future lever – you'll quickly grow tired of the weight of decision and instead end up following some simple rule like "just push" or "just don't push.")

If all your future time slices pull their levers, says the *ex post* deontologist, then to pull *Move Alice* now is to trigger a certain kind of morally disvaluable phenomenon. It is to crush Alice for the sake of someone else or to crush someone else for the sake of Alice. (Say that this phenomenon has an agent-relative choiceworthiness of -50.) If none of your future slices pull their levers, says the *ex post* deontologist, then to pull *Move Alice* now is to open up a couple of different possibilities. Alice might turn out to be upon the bridge, such that pulling the *Move Alice* lever causes her to perish alone on the second track. (Say that this outcome has an agent-relative choiceworthiness of -10.) But Alice might also turn out to be on the tracks (twice as likely) such that pulling the *Move Alice* lever causes her to survive alone on the second track, protected by a 'sand suitcase.' (Say that this outcome has an agent-relative choiceworthiness of 10.) Your choice could thus be represented as follows:[39]

Table 2. Agent-relative choiceworthiness in ***Suitcases (Parallel Tracks)***

	All future time slices pull (0.4)	No future time slices pull, Alice is on track (0.4)	No future time slices pull, Alice is on footbridge (0.2)
Pull *Move Alice*	-50	10	-10
~Pull *Move Alice*	0	0	0

[39] The agent-relative unchoiceworthiness of committing to a bad-kind killing must outweigh (*ex hypothesi*) the agent-relative choiceworthiness involved in saving two lives, or else the *ex post* view would fail to capture the orthodox deontological verdict in **Footbridge**.

Under these toy numbers, at least, the expected choiceworthiness calculations speak in favour of not pushing.[40] Given that you're moderately confident that you'll later go on to pull other levers, you can be moderately confident that pulling *Move Alice* will contribute to an especially disvaluable bad-kind killing, such that this option is unchoiceworthy in comparison with the status quo.

Since *ex post* deontologists can easily resist premise (1) of the Composition Argument, the argument seems to lack dialectical force. It is an argument which *ex post* deontologists need not fear.

3.2 The Deontic Cycling Argument

A second argument seems to improve on the first. Call this the argument from *deontic cycling*. Unfortunately, the argument relies upon an even more complicated case. Thus consider Kowalczyk's:

> **Suitcases III.** As in **Suitcases (Parallel Tracks)**. But this time you don't directly pull the levers. Over several hours, you'll instead have the chance to press a magical button which pulls some number of levers for you. The button can only be pressed once. Press the button at 12:00 and no *Move* levers will be pulled. Press the button at 12:30 and all three *Move* levers will be pulled. Press the button at 1:00 and the *Move Alice* and *Move Blaise* levers will be pulled. Press the button at 1:30 and the *Move Alice* lever will be pulled. Press the button at 2:00 and no *Move* levers will be pulled. Pressing the button at 2:00 (rather than 12:00) is slightly better for whoever ends up surviving – it gives them more time to 'get used to mortality.'[41]

The general idea behind the case is that you can change how many *Move* levers will be pulled, depending upon the time at which you decide to press your button. Kowalczyk

[40] Pushing has expected choiceworthiness of 0.4*-50+0.2*-10+0.4*10=-18, which is lower than the status quo. One relevant complication is that, for evidentialists, pushing now might constitute evidence that you are more likely to push later. I put this complication aside.
[41] Kowalczyk (2022: 19).

shows that, in **Suitcases III**, *ex post* deontology has the following rather shocking implication. You should press the button at 12:00 (moving nobody) rather than pressing the button at 2:00 (moving nobody). This is even though pressing the button at 2:00 (moving nobody) would have been better for the single surviving person. Since this implication is absurd, the thought goes, *ex post* deontology must ultimately be abandoned.

Why does *ex post* deontology yield this implication? Well, suppose it is permissible to press the button at 2:00. That is, suppose it is permissible to trigger *no* levers. (This is, of course, just the sort of claim that the *ex post* deontologist is trying to defend.) It would then be more choiceworthy to push at 1:30, as this is better for the prospects of just one additional person.[42] It would then be more choiceworthy to push at 1:00, as this is better for the prospects of just one additional person. It would then be more choiceworthy to push at 12:30, as this is better for the prospects of just one additional person. But it would then be more choiceworthy to push at 12:00, as pressing at 12:00 rather than at 12:30 ensures that you avoid doing a bad-kind killing of one for the sake of two others. So *ex post* deontologists are committed to the conclusion that it is more choiceworthy to press at 12:00 rather than to push at 2:00:

Increasing choiceworthiness of options, conditional on Option 5 being permissible



Options in ***Suitcases III***

Option 1. Press the button at 12:00, triggering no levers.
Option 2. Press the button at 12:30, triggering all three levers.
Option 3. Press button at 1:00, triggering *Move Alice* and *Move Blaise* levers.
Option 4. Press the button at 1:30, triggering the *Move Alice* lever.
Option 5. Press the button at 2:00, triggering no levers (and minorly benefitting whoever is on the bridge).

[42] The principle at work here, and in the next several steps, is *One Person Ex Ante Pareto*: when your choice affects one and no others, you should do what is best for that single person's prospects.

But this is a seemingly absurd result: it is seemingly absurd that you should choose to move no levers and leave one person poorly off rather than to move no levers and leave that same person better off.

Having spent so long on the *Composition Argument*, I don't think we need to spend as long on the *Deontic Cycling* argument. In short, for quite similar reasons, it's not obvious to me that the deontic cycling argument gives us any additional dialectical traction against *ex post* deontologists.

Ex post deontologists were committed all along, we knew, to the idea that there is something especially morally awful about destroying someone's body for the sake of benefitting others. We knew as much from the original **Suitcases**. But this means that *ex post* deontologists need not accept the above ranking of options in terms of comparative choiceworthiness. In particular, the *ex post* deontologist would be within their rights to judge that, within the above sequence of options, Option 3 (say) is *less* choiceworthy than Option 4. For instance: it's true that Option 3 improves the prospects of one additional person when compared with Option 4. Nonetheless, there is another sense in which *ex post* deontologists are going to be deeply averse to Option 3 when comparing it with Option 4. Option 3, unlike Option 4, carries a risk of entangling myself in a causal chain in which I destroy one person's body *as a means of saving someone else's body*. It carries a risk of taking Alice's body and making it into a train-stop for Blaise's body, or taking Blaise's body and making it into a train-stop for Alice's body. This is a risk which simply would not have been present, had the agent contented herself with Option 3: had she contented herself with *merely* moving Alice to the second track.

Kowalczyk suggests that, were the *ex post* deontologist to offer this species of response, it would be deeply unappealing: it would be an instance of *causal fetishism*, caring more about the details of the causal structure of the case itself than about promoting the

interests of the particular persons who stand to be affected by our choice.[43] I am not so convinced by this reply, though. At least, I am not convinced if Kowalczyk's aim was to develop an argument which gave us additional and independent traction against *ex post* deontologists in the first place.

After all, in general, a deontologist who believes in agent-relative constraints against the infliction of harm is likely to be unmoved by the accusation of causal fetishism. Such deontologists were already perfectly upfront about their belief that there are certain harmful causal chains in which we ought not entangle ourselves, even if refraining from entangling ourselves in those causal chains comes at a cost. To take Hanser's familiar case, suppose that I and several other assassins are lined up to stab the same target.[44] The standard deontological judgement, I take it, is that I should view it as especially unchoiceworthy for *me* to plunge in the blade, such that *I* become the killer, even if the same outcome is overdetermined whichever choice I end up making. Many deontologists would wish to maintain this verdict, I suspect, even if my assassination would be the most pleasant for the victim's welfare by some small margin – even if I could accomplish the assassination in a slightly less painful way, say.

If deontologists already affirmed that we must not 'causally entangle' ourselves in bad-kind killings in cases like Hanser's, even when this makes no difference (or even a *worse* difference) for victims, I take it that it is a very straightforward move for deontologists to resist the *Deontic Cycling* argument by denying that Option 3 is better than Option 4. They can do so precisely on the grounds that to perform Option 3 is to risk entangling myself in a bad-kind killing, but that to perform Option 4 is not.

[43] Kowalczyk (2022: 21).
[44] Hanser (1995).

Of course, those who endorse (say) a purely impartial conception such as act utilitarianism might find this a rather absurd position to take. They might think that, clearly, it makes no difference whoever performs the assassination in Hanser's case. Nonetheless, insofar as the target of the argument is indeed the sort of *ex post* deontologist who intuits there to be something especially wrongful about entangling oneself in a pushing in **Suitcases**, I don't think the argument can be seen as holding additional dialectical force against the *ex post* deontologist.

3.3 Explaining Away the Destruction Intuition

In my view, then, the most recent arguments against *ex post* deontology don't do much to move the dial. In particular, they don't hold much dialectical force if one is in fact debating with an *ex post* deontologist in the first place.

But of course, as I've already acknowledged, my intuitive sympathies lie with *ex ante* deontology. For my part, I think it *may* be possible to develop a dialectically forceful argument in favour of *ex ante* rather than *ex post* deontology. My own sense, however, is that the argument would have to come from a rather different direction. The manoeuvre goes as follows.

As we've already seen, the *ex post* deontologist intuits that it is an especially morally awful phenomenon to destroy one person's body as a means of saving another. It was this intuition – call it the 'destruction intuition' – which explained why *ex post* deontologists were so unhappy with the idea of pushing in the original **Suitcases**.

What explains or underpins the destruction intuition? Frances Kamm once attempted to support the intuition by claiming that to push, in cases like **Suitcases**, would be to fail to *uphold a certain conception of a person as inviolably valuable*. It would be to understand

persons, *qua* the objects of interpersonal morality, as a kind of tool or implement rather than an inviolable end which ought to be treated with dignity. Thus Kamm writes:

> It is more important that *the concept of the person expressed by a constraining right* … be true of people inalienably, and that events accord with the existence of such a right, than it is to maximize chances of survival. This is because while more people may be harmed, and so the chances of each of us being harmed may be greater, the conception of each person that is endorsed involves a high degree of inviolability. We may all lead harder lives, but our dignity is greater.[45]

Perhaps this passage does not illuminate things much further, however: perhaps it is no more helpful than saying, as I did a moment ago, that *ex post* deontologists simply intuit that there is something especially morally awful in destroying someone's body as a means of saving another person's life within **Suitcases**.

As with any popular and widespread intuition, however, it's possible to ask whether the destruction intuition reported by *ex post* deontologists in **Suitcases** is something which can in fact be explained away. There are at least two reasons why I think the destruction intuition in **Suitcases** should be treated with suspicion, even if it turns out to be popular.

The first reason is that, although I do find myself with a very strong 'destruction intuition' in independent cases, these are cases in which independent factors appear best suited to explaining its presence. Upon closer examination, these independent factors turn out to be altogether absent from **Suitcases**.

For instance, I do find myself struck very strongly by the destruction intuition in the independent case where a single identified person voluntarily asks us to kill them, take their organs, and save someone else. In such a case I have the strong intuition that it would be wrong to destroy someone's body for the sake of others, even if everyone

[45] Kamm (1996: 294).

involved is willing for us to do so. But insofar as the destruction intuition emerges in this independent case, it seems that it could be explained and justified by independent factors. For instance, one (admittedly controversial) possibility is that the patient in this independent case holds something of a duty to self, where this duty to self might constrain, somehow, the ways in which they may self-govern.[46] Perhaps this duty implies that the patient cannot simply elect to self obliterate, even if doing so would be altruistic; alternatively, perhaps it implies that although the agent *herself* holds this power, she is unable to gift this power to others. Such independent factors could explain and legitimise the destruction intuition in independent cases. But it would be a stretch, I think, to say that they could explain and legitimise the destruction intuition in a case like **Suitcases**. It is implausible that, in consenting for the suitcase to be pushed in **Suitcases**, any given patient would neglect a self-directed duty of preservation.

The second reason for hesitancy goes as follows. I've already suggested that there are distinctive consent-based reasons against pushing in a case like **Suitcases (One Way Fabric)** – that is, in cases in which some act φ is in the expected interests of all (according to your best subjective probabilities) but in which certain patients already know where they are, such that performing φ would indeed involve harming someone against their express wishes. But in practice, it is enormously difficult to distinguish these cases from the more 'pure' type of case presented by **Suitcases**. We should not be surprised, then, if we often have a negative intuitive reaction to policies which involve destroying someone's body for the sake of everyone's expected interests. It is often hard to pinpoint whether the case in question belongs to the former category or the latter.

[46] For very recent discussion of self-wronging see Muñoz & Baron-Schmitt (2024).

For instance, let's suppose that the lever in **Suitcases** must be pulled manually over a period of two seconds – as you pull the lever, the suitcase upon the footbridge is gradually tilted off the edge. (I assume that this is how things like levers often work in the real world.) The moment that this is the nature of the case, then the case becomes morally analogous to **Suitcases (One Way Fabric):** if I were to continue pressing the lever all the way down, I would be killing someone who now (feeling that it is *their* body beginning to tilt at the pull of the lever) knows where they are. Such a person does not, presumably, give me consent to continue imposing myself upon their body in such a way. In this way I would have distinctive consent-based reasons against pushing: it would be *right*, to some degree, to have a negative intuition reaction to pushing.

Such a negative intuitive reaction would be warranted. It would be a mistake, however, to infer from this result that a negative reaction is warranted in the most abstracted and idealised version of **Suitcases**, as the *ex post* deontologist seems to do – that is, in a case in which the agent destroys a body for the sake of others, where *everyone*, until the very moment of choice, believes that the destruction is in their expected interests.

To respond to the *ex post* deontologist in this way is to explore a rather different avenue. It is to ask whether there are plausible methods for explaining away the destruction intuition which *ex post* deontologists report in **Suitcases** in the first place. The above methods strike me as having promise. If we could indeed explain away the intuitions which *ex post* deontologists report in cases like **Suitcases**, then perhaps there is a plausible way of arguing against the *ex post* deontologist in fully theory-neutral terms: that is, a way of arguing against *ex post* deontologists which even deontologists can agree with.

4. Conclusion

I first argued that we may only harm out of kindness: we may only harm when doing so is in the expected interests of each and every person. This is the attractive middle-ground provided by *ex ante* deontology. I showed, in particular, that the diachronic objection to *ex ante* deontology is unsuccessful: sophisticated models of *ex ante* deontology deliver the correct verdict that we may never lethally and non-consensually impose ourselves upon the bodies of others.

I then considered whether we ought to prefer an *ex ante* deontological model or an *ex post* deontological model. I argued that, although recent arguments against *ex post* deontology have tended to lack dialectical force, there may indeed be a promising theory-neutral avenue for demonstrating the superiority of the *ex ante* model. The trick is to try and explain away the intuitions which underpinned the *ex post* deontological model in the first place.

References

Dougherty, T. (2014). 'Fickle Consent', *Philosophical Studies*, 167(1): 25–40.

Dworkin, G. (1972) 'Paternalism', *The Monist*, 56(1): 64–84.

Foot, P. (1967) 'The Problem of Abortion and the Doctrine of Double Effect', *Oxford Review*, 5: 5–15.

Frankfurt, G. (1971) 'Freedom of the Will and the Concept of a Person', *The Journal of Philosophy*, 68(1): 5–20.

Gauthier, D. (1987) *Morals By Agreement*. Oxford University Press.

Glover, J. (1977) *Causing Death and Saving Lives*. Penguin.

Gustafsson, J. E. (2022) *Money-Pump Arguments*. Cambridge University Press.

Hanser, M. (1995) 'Why are Killing and Letting Die Wrong?', *Philosophy & Public Affairs*, 24(3): 175–201.

Hare, C. (2013) *The Limits of Kindness*. Oxford University Press.

——— (2016) 'Should We Wish Well to All?', *Philosophical Review*, 125(4): 451–472.

Harris, J. (1975) 'The Survival Lottery', *Philosophy*, 50(1): 81–87.

Hey, J. D. & Lotito, F. (2009). 'Naive, Resolute or Sophisticated? A Study of Dynamic Decision Making', *Journal of Risk and Uncertainty*, 38(1): 1–25.

Hooker, B. (2015) 'The Elements of Well-Being', *Journal of Practical Ethics*, 3(1): 15–35.

Kagan, S. (2011) 'Do I Make a Difference?', *Philosophy & Public Affairs*, 39(2): 105–141.

Kamm, F. (1996) *Morality, Mortality: Volume II: Rights, Duties, and Status*. Oxford University Press.

Kowalczyk, K. (2022). 'People in Suitcases', *Journal of Moral Philosophy*, 20(1–2): 3–30.

Liberto, H. (2017) 'The Problem With Sexual Promises', *Ethics*, 127(2):383–414.

Machina, M. (1989) 'Dynamic Consistency and Non-Expected Utility Models of Choice Under Uncertainty', *Journal of Economic Literature*, 27(4): 1622–1668.

Maher, P. (1992) 'Symptomatic Acts and the Value of Evidence in Causal Decision Theory', *Philosophy of Science*, 57(3): 479–498.

McClennan, E. F. (1990) *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge University Press.

Muñoz, D. & Baron-Schmitt, N. (2024). 'Wronging Oneself', *Journal of Philosophy*, 121(4): 181–207.

Murphy, L. (2014) 'The Practice of Promise and Contract', in G. Klass, G. Letsas, & P. Saprai (eds), *Philosophical Foundations of Contract Law*. Oxford University Press.

Oberdiek, J. (2017) *Imposing Risk: A Normative Framework*. Oxford University Press.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Setiya, K. (2020) 'Ignorance, Beneficence, and Rights', *Journal of Moral Philosophy*, 17(1): 56–74.

Stark, C. (2000). 'Hypothetical Consent and Justification', *The Journal of Philosophy*, 97(6): 313–334.

Strotz, R. (1955) 'Myopia and Inconsistency in Dynamic Utility Maximisation', *The Review of Economic Studies*, 23(3): 165–180.

Sun, A. (2025) 'Can Consent Be Irrevocable?', *Philosophers' Imprint*, 25(38): 1–17.

Thomson, J. (1985) 'The Trolley Problem', *The Yale Law Journal*, 94(6): 1395–1415.

——— (1990) *The Realm of Rights*. Harvard University Press.

Unruh, C. F. (2023) 'The Constraint Against Doing Harm and Long-Term Consequences', *Journal of Moral Philosophy*, 20(3–4): 290–310.

Voorhoeve, A. & Fleurbaey, M. (2013) 'Decide as You Would With Full Information! An Argument Against *Ex Ante* Pareto', in N. Eyal, S. Hurst, O. Norheim, and D. Wikler (eds), *Inequalities in Health: Concepts, Measures, and Ethics*. Oxford University Press.

6.

A Theory of Ex Ante Designation

Abstract. There are different ways in which agents might designate patients at the *ex ante* stage of moral decision making. It turns out, however, that changing the mode of designation can change the deontic status of one's options under various *ex ante* moral standards. This is bad news. In this paper, I solve the problem by developing a *normative theory of ex ante designation*: a theory which specifies how agents ought to designate patients at the *ex ante* stage. The theory I develop is multicriterial. Agents ought to make use of designators which facilitate singular thoughts about patients. And agents ought to make use of those designators which are maximally loaded with identity-relevant information. In principle, these two criteria can conflict in a way which does not admit of an easy resolution; this, I suggest, is a theoretical consequence we ought to embrace.

1. Introduction

This article develops a *normative theory of ex ante designation:* a normative theory, that is, which specifies how agents ought to name patients at the *ex ante* stage of moral decision making. Why are we in need of such a theory? It turns out, in brief, that there are different ways in which agents might designate patients at the *ex ante* stage. Changing the mode of designation can change the prospects held by individual patients. Changing the mode of designation can thus change the deontic status of one's options under various *ex ante* moral principles. Unless we develop a theory which specifies *how* agents ought to name patients,

then these various *ex ante* moral principles are in trouble: they fail to qualify as complete and action-guiding normative standards.[1]

The theory of designation I develop is multicriterial: it says that there exist multiple independent criteria which agents ought to satisfy in their attempts to designate patients. According to one such criterion, agents have reason to use those designators under which they can think of patients 'singularly.' According to another such criterion, agents have reason to use those designators which are maximally 'informationally loaded.'

Each of these criteria is supported by an independent rationale which demonstrates its normative importance. The singular thought criterion is important insofar as it allows agents to uphold the separateness of persons, and insofar as it allows agents to be directly motivated by the claims of those they attempt to aid. The informational criterion is important as it facilitates an aspect of interpersonal morality which we should view as valuable – an aspect of interpersonal morality which I call *recognition*.

These two criteria will sometimes conflict in a manner which does not admit of an easy resolution. This, I suggest, is an upshot we ought to accept. In cases of criterial conflict there may simply be no determinate answer about how agents ought to designate those persons who stand to be affected by their choice. It follows that, in such cases, there may be no determinate answer about the recommendations delivered by *ex ante* moral principles. If this is a dark cloud then a silver lining, as we will see, is that the two criteria speak in unison in an extremely wide range of practically relevant cases.

[1] Of course, I defended certain *ex ante* principles in previous chapters. In doing so I previously put the problem of multiple designations to one side. Now, I treat the problem in full. The problem has been diagnosed with greatest depth by Mahtani (2017; 2021). Recent discussions of this and other closely related puzzles are found in Hare (2016), Setiya (2020; 2023), and Gustafsson & Kowalczyk (forthcoming).

2. *Ex Ante* Principles and the Problem of Multiple Designations

To begin with, let me broadly sketch the family of moral principles in which I am interested. Across contemporary normative ethics, when discussing some moral principle *P*, a distinction is regularly drawn between the *ex post* construal of *P* and the *ex ante* construal of *P*. In these pages it is the latter group of principles in which I am interested.[2]

If we wanted to draw the distinction between *ex post* and *ex ante* very broadly, we might say the following. The *ex post* construal of a given moral principle *P* typically makes judgements of permissibility *vis-à-vis* some act φ by citing, somehow, structural features of the outcomes $o_1 - o_n$ which might eventually follow φ's performance.[3] By contrast, the *ex ante* construal of a given moral principle *P* typically makes judgements of permissibility *vis-à-vis* some act φ by foregrounding, somehow, features of the *prospects* that φ gives to each individual person $i_1 - i_n$ who stands to be affected by the performance of φ. To find an individual i's *prospects* under an act φ, one must take the probability-weighted sum of i's well-being across each of the possible outcomes $o_1 - o_n$ that might follow φ's performance.[4]

Theorists sympathetic to *ex ante* moral standards, such as myself, face a severe challenge. In articulating and responding to this challenge, I'll focus upon a single concrete example of an *ex ante* normative principle. Following Anna Mahtani, I'll take as my focus the *ex ante Pareto Principle*, a canonical principle of social choice and welfare

---

[2] On *ex post* versus *ex ante* prioritarianism see McCarthy (2006; 2008). On *ex post* versus *ex ante* non-aggregation, see Frick (2015), Horton (2017), Rüger (2018), and Walen (2020). On *ex post* versus *ex ante* equality see Temkin (2009). On *ex post* versus *ex ante* rank-weighted utilitarianism, see Nebel (2020). For a treatment of the literature on *ex post* versus *ex ante* deontology, see my previous Chapter 5.

[3] By necessity this statement is broad: see my Chapter 4 for a discussion of *ex post* partial aggregation, for instance, according to which the strength of the strongest claim in favour of a given act φ (associated with potential outcomes $o_1 - o_n$) is given by the weighted average of the strongest single claim satisfied by each of $o_1 - o_n$.

[4] One could allow, here, the probabilities to be transformed by some suitable function representing permissible risk aversion.

economics according to which an act φ is impermissible if an alternative act ψ would provide each individual with just as good prospects and at least one individual with better prospects. Say, of any act that fails this condition, that it is *ex ante Pareto impermissible*.[5]

Here is the challenge. Suppose you are a proponent of *ex ante* Pareto. It turns out that, when allocating prospects to individuals at the moment of choice, there are different ways in which you might decide to designate those individuals. Different modes of designation seem to change the prospects that those individuals have, and thus, seem to change whether or not your behaviour qualifies as impermissible from the standpoint of *ex ante* Pareto. Call this the *problem of multiple designations*. The problem of multiple designations is most easily brought out in the following sort of case:

> **Suitcases.** Five persons are about to be killed by an oncoming trolley. You can save them by pressing the *Push* button, which will shunt a sixth person off an overhead footbridge, fatally mangling that sixth person but stopping the trolley prematurely. The names of the six persons are Alice, Blaise, Charlie, Dora, Emilio, and Frederica. The catch? Each of these six persons are zipped shut in a different suitcase, and you haven't a clue as to which person is in which of the suitcases. Ultimately you decide to push the button.[6]

You might think, upon reflection, that it is in fact unclear how a standard like *ex ante* Pareto applies to **Suitcases**. Here is why.

In **Suitcases** one could, perhaps, allocate prospects to a set of persons designated as *Alice*, *Blaise*, *Charlie*, *Dora*, *Emilio*, and *Frederica*. When persons are designated in such a way, and when prospects are calculated according to this mode of 'mental presentation',

---

[5] Mahtani (2017; 2021). For some of the controversies surrounding *ex ante* Pareto, see Fleurbaey and Voorhoeve (2013) and Adler (2011: Chapter 7). I follow Mahtani's (2017: 307) usage that the *ex ante* Pareto principle is a constraint on *actions*: a constraint which specifies what acts you should or should not perform. Throughout, I use the language of 'permissibility' and 'impermissibility', but the central puzzle would still have emerged had we stuck with mere 'shoulds' and 'should nots.'

[6] Hare (2016), with previous analogues in Harris (1975), Glover (1977), and Kamm (1996). I discussed the **Suitcases** example at length in my Chapter 5. There, as mentioned, I simply took for granted that pushing was to be construed as *ex ante* optimal, temporarily putting aside the puzzle of *ex ante* designation.

then *Push* will improve each individual's prospects. To ~*Push*, then, would be to violate *ex ante* Pareto:[7]

Table 1. Prospects presented by proper names

	Push	~*Push*
Alice	1/6 die, 5/6 live	5/6 die, 1/6 live
Blaise	1/6 die, 5/6 live	5/6 die, 1/6 live
Charlie	1/6 die, 5/6 live	5/6 die, 1/6 live
…	…	…

Clearly, though, this is not the only way of designating – of 'mentally presenting' – the set of persons who stand to be affected by your choice. For instance: you know that a given individual, whatever their proper name, is in the suitcase upon the bridge. It thus seems perfectly reasonable to calculate prospects for *The Person In A Suitcase On The Bridge*. If you do so, your decision to press the *Push* button will result in worsened prospects for *The Person In A Suitcase On The Bridge*. To ~*Push*, then, would no longer violate *ex ante* Pareto:

Table 2. Persons presented by spatial location

	Push	~*Push*
The Person In A Suitcase On The Bridge	Certain death	Certain life
The Closest Person On The Tracks	Certain life	Certain death
The Second-Closest Person On The Tracks	Certain life	Certain death
…	…	…

Similarly, suppose there exists a small hole in the suitcase on the bridge.[8] You can see through the hole that the person in the suitcase on the bridge has brown hair. *Mutatis mutandis* for those persons stuck in suitcases on the track. (You remain clueless, unfortunately, about the physical appearance of the individuals who go by the names Alice,

[7] Throughout, I assume that the relevant probabilities to be used in the calculation of prospects are the epistemically rational ones: the probabilities that the decision-maker herself ought to hold in light of her available evidence. For discussion of some of the relevant complications, see Oberdiek (2017: Chapter 1).
[8] Hare (2016: 467).

Blaise, Charlie … and so on.) It seems perfectly reasonable to now calculate prospects for *The Person With Brown Hair*. If you do so, your decision to press the *Push* button will result in worsened prospects for *The Person With Brown Hair*. To *~Push*, then, would no longer violate *ex ante* Pareto:

Table 3. Persons presented by physical characteristics

	PUSH	~PUSH
The Person With Brown Hair	Certain death	Certain life
The Person With Black Hair	Certain life	Certain death
The Person With Red Hair	Certain life	Certain death
…	…	…

To give one final example, suppose you know that the person in the suitcase on the bridge commonly goes by the nickname *Ghost Rider*, where *Ghost Rider* enjoys the occasional game of football. You also know the nicknames used by those persons stuck in suitcases on the tracks. (You remain clueless, unfortunately, about whether or not the nickname *Ghost Rider* is the one used by Alice, Blaise, Charlie … and so on.) As before, it seems perfectly reasonable to calculate prospects for *Ghost Rider*. But if you do so, your decision to *Push* will result in worsened prospects for *Ghost Rider*. To *~Push*, then, would no longer violate *ex ante* Pareto:

Table 4. Persons presented by nickname

	Push	*~Push*
Ghost Rider	Certain death	Certain life
Green Lantern	Certain life	Certain death
Flash	Certain life	Certain death
…	…	…

The problem in its most general form, then, is that changing between various and seemingly reasonable modes of designation will change whether or not one's action is permitted according to *ex ante* principles such as *ex ante* Pareto. The problem, as Mahtani notes, is one with Fregean roots: just as it is possible to know distinctive features of the

morning star and the evening star without knowing whether those terms co-refer to a numerically identical celestial object, it is possible to know distinctive features of Alice and *Ghost Rider*, without knowing whether those terms too co-refer to a numerically identical patient.[9]

3. A Theory of *Ex Ante* Designation

One might take the problem of multiple designations to simply show that *ex ante* standards are ill-defined, and that, as a consequence, moral analyses of risky decision making ought to always operate in a manner which is *ex post*. My goal here, however, is more constructive than this. It is to investigate whether we can develop a *normative theory of ex ante designation*: a theory which would specify, in virtue of fully general criteria, how agents ought to designate patients at the *ex ante* stage.

The theory I develop is multicriterial. It says, in other words, that there exist multiple independent criteria that agents ought to satisfy in their attempts to designate patients. In what follows I'll present each criterion in turn, explain how that criterion applies to cases like **Suitcases**, and show how that criterion can be supported by an underlying normative rationale. Having presented the criteria, I'll consider what we ought to say with respect to cases of criterial conflict.

[9] Mahtani (2017). Mahtani's own proposal is the revisionist one that we should understand an act as satisfying *ex ante* Pareto if it satisfies *ex ante* Pareto under each candidate mode of designation. As will become apparent, I think this proposal gives up too quickly on the idea that there is often a uniquely appropriate way of designating patients. Nonetheless, in §5, I consider certain cases of criterial conflict where Mahtani's proposal might strike us as especially useful.

3.1 The Singular Thought Criterion

Imagine that you can use two sets of designators ($D_1 - D_n$ and $E_1 - E_n$) to label the individuals $i_1 - i_n$ who stand to be affected by your choice. Which set of designators ought you use as the basis for the allocation of prospects?

One possibility goes as follows. Insofar as it is possible, you ought to label the individuals $i_1 - i_n$ using those designators which will allow you to *actually think thoughts* about those individuals who stand to be affected by your choice. Insofar as the set of designators $D_1 - D_n$ rather than $E_1 - E_n$ allows you to "actually think thoughts" about the individuals who stand to be affected by your choice, that gives you a reason in favour of employing $D_1 - D_n$ rather than $E_1 - E_n$.

The phrase "actually think thoughts" is, of course, terribly vague. We are in want of a way of spelling out the details. One possibility, I believe, would be to draw upon a long-established tradition in philosophy of mind and language which draws a conceptual distinction between so-called *singular thoughts* and *general thoughts*. Here, then, would be the current proposal. Insofar as it is possible, at the *ex ante* stage, I ought to employ those sets of designators which allow me to *think singular thoughts* about the individuals who stand to be affected by my actions:

> **The Singular Thought Criterion.** Insofar as it is possible, at the *ex ante* stage, agents ought to employ those designators which facilitate singular thought about the individuals affected by their actions.[10]

There remains disagreement amongst philosophers of language and mind, of course, as to the exact nature of the singular versus general distinction in the first place. Speaking very broadly, we might draw the distinction as follows. Some thoughts aim to be about *a thing*

[10] Something close to the singular thought criterion is briefly considered and rejected by Hare (2016: 469). See, too, remarks at the conclusion of Setiya (2020: 71–73).

in particular; these thoughts are the singular ones. Some thoughts do not aim to be about a thing in particular; these thoughts are the general ones.[11]

The relevant distinction is most easily motivated in concrete examples. Suppose I find myself idly pondering who, a hundred years from now, will be the tallest person in Spain.[12] I mentally baptise such a person *The Tallest Person In Future Spain*. And I proceed to have the thought: "I hope *The Tallest Person In Future Spain* has a good life." Cases such as this – cases in which arbitrary descriptive names are introduced to refer to entities with which I am not in any way acquainted – are canonical cases in which philosophers of language deny that the thought in question is singular.[13] My thought about *The Tallest Person In Future Spain* fails to qualify as singular, presumably, because as things seem to me, at the moment of thinking, there are or will be countless individuals out there in the world who might eventually turn out to be *The Tallest Person In Future Spain;* I dwell upon none of those individuals *in particular* insofar as I think about *The Tallest Person In Future Spain*.

By contrast, suppose I am out hiking when I encounter a stranger coming back down the path. I find myself struck by the following thought: "That stranger's yellow raincoat is blindingly bright." Cases such as this – cases in which I entertain a thought directed towards an object *o* with which I am directly acquainted – are canonical cases in which philosophers of language typically *grant* the possibility of singular thought. My thought about the yellow raincoat qualifies as singular, presumably, insofar as the thought seems to play the cognitive role of aiming towards a particular object – the yellow raincoat which has just entered my field of view and which I observe to be keeping the stranger dry.

---

[11] The distinction isn't to be confused with the distinction between singular and plural thoughts; more on this point shortly. I follow Crane's (2011) terminology that singular thoughts are ones which *aim* or *purport* to be about a particular object.

[12] This is derivative of Kaplan's (1989) *Newman 1*, an introduced descriptive name used in order to pick out the first person born in the 22nd century.

[13] See, for instance, Jeshion's (2010: 114–116) extended list of non-singular examples.

In such cases, it does indeed seem that an intuitively sensible conceptual distinction can be drawn between singular versus general thought. The *orthodox analysis* of singular thought infers, from such cases, two general conditions that must be satisfied in order for singular thoughts to be realised.[14] The first condition says that, in order for me to think singularly of some object *o*, I must be in the special epistemic relationship of *acquaintance* with *o*. The second condition says that, in order for me to think singularly of some object *o*, my singular thought about *o* must metaphysically depend, somehow, upon the existence of *o* itself. It is in virtue of these two conditions being satisfied, the suggestion goes, that my thought about the stranger's yellow raincoat can aptly be characterised as singular.

This orthodox analysis of singular thought naturally inspires the further question of what, exactly, it takes for me to become acquainted with some object *o* (in the sense relevant for sustaining singular thought about *o*). One possibility is simply to say that acquaintance with *o* requires past or present perceptual contact, of some form or another, with *o*. I proceed, in the following sections, under the ecumenical assumption that perceptual contact with *o* is at least *one* means via which I can become acquainted with *o* and thus begin to think singularly about *o;* this is not to deny that singular thought could come about via other mechanisms of acquaintance too.[15]

[14] I borrow the term 'orthodox' from Crane (2011), and in doing so, don't mean to deny that the distinction isn't hotly contested. For a canonical statement of these two components see Armstrong & Stanley (2011).
[15] As Jeshion (2010: 10), notes, many theorists are tempted by a broader conception of acquaintance: what we might call the *Standard Standard Of Acquaintance*, according to which, to be acquainted with *o* (in the sense requisite for singular thought about *o*) I must have perception of *o*, memory of *o*, or access to communication chains concerning *o* (which themselves end in someone's perception of *o*). Under the *Standard Standard*, acquaintance with *o* always presupposes that at least someone within one's linguistic community has, at some stage, had perceptual contact with *o*.

3.2 Applying the Singular Thought Criterion

Consider how the singular thought criterion might apply to **Suitcases**. Suppose, by way of spelling out the details, that I have previously become acquainted with (have previously seen, perhaps) each of the six individual persons I know to be Alice, Blaise, Charlie, and so on.

Consider my thought: "I hope *Alice* is well." Is this thought a singular one? Quite plausibly, the answer is *yes*. For my thought seems to be aimed towards a particular object – namely, towards the Alice with whom I have previously interacted. *Alice* thus qualifies as a legitimate *ex ante* designator. I am permitted to calculate prospects for *Alice*. (*Mutatis mutandis* for Blaise, Charlie … and so on.)

Consider, by contrast, the thought: "I hope *The Person In The Suitcase On The Bridge* is well." Is such a thought singular? Quite plausibly, the answer to this question is *no*. For my thought about *The Person In The Suitcase On The Bridge* does not seem to be aimed towards a particular object. There is of course a *sense* in which that thought is aimed towards a particular object: given that the case is a deterministic one, some given individual must certainly satisfy the definite description of *The Person In The Suitcase On The Bridge*. But at the *ex ante* stage, as things seem to me *qua* decision maker, there are many candidate individuals who might equally satisfy such a description – just as there are many candidate individuals who (as things seem to me) might satisfy the description of *The Tallest Person In Future Spain*. For this reason, it seems that the thought "I hope *The Person In A Suitcase On The Bridge* is well" fails to qualify as singular, just as the thought "I hope *The Tallest Person In Future Spain* has a good life" fails to qualify as singular.[16]

[16] There is an important but nonetheless finnicky complication here. If I am acquainted with Alice (i.e., have seen her before), and it just so turns out that Alice *is The Person In The Suitcase On the Bridge*, doesn't

What we end up endorsing in this initial contest, then, is the use of proper names – Alice, Blaise, Charlie, and so on – as the basis for the allocation of prospects, rather than the use of designators such as *The Person In The Suitcase On The Bridge*. We receive the verdict, at least in this initial contest, that ~*Push* is prohibited by the *ex ante* Pareto principle. The alternative act, *Push*, would have benefitted the prospects of all and harmed the prospects of none.

One might resist this initial conclusion, pushing back in favour of designators such as *The Person In The Suitcase On The Bridge*. The pushback would go as follows. It's true, admittedly, that I can't see the actual (say) skin pores of *The Person In The Suitcase On The Bridge*. But one might nevertheless think that, in **Suitcases**, I have had perceptual contact with *The Person In The Suitcase On The Bridge* (so presented) sufficient for generating singular thought about that individual (so presented). In the case as described, for instance, I can still observe the bumps in the suitcase made by the distinctive shape of that individual's body. Perhaps I can poke the side of their suitcase and hear a fleshy *squish*. And so on. The current pushback would maintain that indirect sensory perceptions such as these should be sufficient for acquaintance with (and thus singular thought about) *The Person In The Suitcase On The Bridge*, since they are still perceptual experiences caused by

---

this entail that I am acquainted with (and can thus think singularly about) *The Person In The Suitcase On The Bridge?* Not quite. For to think singularly about some object *o* in fact requires that some further condition is satisfied: a condition which Jeshion (2010: 109) helpfully labels the *Acquaintance Basis* condition. Roughly, for me to think singularly about some object *o*, my thought about *o* must be based upon my acquaintance with *o*. If I am clueless as to whether Alice *is The Person In The Suitcase On The Bridge*, then my thought about *The Person In The Suitcase On The Bridge* cannot be based upon my acquaintance with Alice. Thus, my thought about *The Person In The Suitcase On The Bridge* fails to qualify as singular. Such an example is complex, but the general idea behind the *Acquaintance Basis* can be more easily grasped in the following sort of example. Suppose I have in fact been acquainted with every bird on planet earth, but I do not know which of them can fly highest. If that is the case, then my thought "I hope *The Highest Flying Bird On Earth* is well" plausibly fails to qualify as singular. For although I *certainly am* acquainted with the bird which will inevitably be picked out by such a label, I do not know *which* bird it is that is inevitably picked out by such a label. So my acquaintance with such a bird cannot serve as the basis for singular thought about *The Highest Flying Bird On Earth*.

– and in some broad sense, *distinctively about* – the individual who falls under the label *The Person In The Suitcase On The Bridge.*

I don't much mind if someone wants to construe 'perceptual contact' in very broad and liberal terms such as these. Nonetheless, importantly, even this broad and liberal construal of perceptual contact would still leave us with the following sort of result. Suppose that the victim in question is not in a suitcase on the bridge, but rather, in a suitcase hidden somewhere in one of several nearby villages. You do not know which village. But you can, as before, push a button which will destroy that person's body in order to save the lives of five en-suitcased others. (As before: you do not know whether that single imperilled individual is called Alice, Blaise, Charlie … and so on).

In such a modified version of the case, you cannot directly or indirectly perceive *The Person In A Suitcase In A Nearby Village.* You cannot even ostend – point – at the general direction of such a person and say "There! *That* is the person who is about to die!" So, even if we endorse a fairly broad and liberal view as to what sorts of perceptions necessitate acquaintance, we will still have to deny that you can have singular thoughts concerning *The Person In A Suitcase In A Nearby Village.* Such an individual so-designated will be incapable of featuring in singular thoughts. It will be impermissible, according to the singular thought criterion, to use this designator in allocating prospects.

3.3 Justifying the Singular Thought Criterion

I've outlined the singular thought criterion and provided an initial sketch of how it might be applied to cases such as **Suitcases.** But the natural question remains: why believe in the singular thought criterion in the first place? Why believe that it is good, insofar as it is possible, for agents to designate patients by making use of designators which facilitate singular (rather than merely general) thought about patients?

The answer to this question is that I believe in:

> *The Concern Claim.* I show a special and normatively important kind of concern to the individuals $i_1 - i_n$ affected by my choice insofar as I choose to employ a mode of designation which facilitates singular thought about those individuals.

Of course, to justify the singular thought criterion by citing the *Concern Claim* is to merely kick the bucket down the road. We are still in want of an underlying justification of the *Concern Claim* itself. Why does using designators conducive to singular thoughts facilitate a normatively important kind of concern? What does this normatively important kind of concern amount to?

There are two broad ways of buttressing the *Concern Claim* with an underlying justification. Here is the first. Making use of singular thoughts, in one's deliberations, seems a distinctive way of upholding the *separateness of persons*. The phrase 'separateness of persons' is of course polysemous. But in very general terms, those who emphasise the separateness of persons typically wish to emphasise that our moral obligations are held towards concrete, flesh-and-blood individuals, each a rational and distinctive locus of personhood, rather than towards generic kinds or collections of entities. Non-consequentialists have, in this vein, long objected that consequentialist standards of rightness violate the separateness of persons by allowing for interpersonal trade-offs of a problematic kind: by allowing for us to take n from A in order to give $2n$ to B, as if A and B were both to be viewed as part of some singular 'clonal colony.'[17]

Plausibly, there is a deep sense in which conforming to the singular thought criterion is prerequisite if we are to respect and uphold the separateness of persons. Suppose, in the above case, that I am acquainted with Alice, Blaise, Charlie … and so on. The singular

[17] This phrase is from Fischer (2024), who provides a detailed survey of existing interpretations. For influential articulations of the interpersonal trade-off concern, see Nagel (1970: 138), Rawls (1971: 24–27), and Nozick (1974: 32–33).

thought criterion would thus prohibit me from deliberating about my moral obligations towards *The Person In A Suitcase In A Nearby Village*. This requirement would demand, in some important sense, that I deliberate about the *particular separate individuals* who stand to be affected by my choice. It would prohibit me, in short, from thinking of myself as holding moral obligations towards a certain kind of mental disjunction of persons, the 'thing which has a chance of being Alice *or* a chance of being Blaise *or* of being Charlie … and so on.' Indeed, insofar as there is a genuine conceptual distinction to be drawn between singular versus fully general thoughts, as many have believed, then there *must* be such a sense in which the singular thought criterion allows moral agents to respect and uphold the separateness of persons. For the singular thought criterion simply says that, insofar as it *is* possible to deliberate using thoughts about separate moral patients who stand to be affected by one's choice in the first place, those thoughts are the ones we ought to employ in our moral deliberations.[18]

The second broad way of buttressing the *Concern Claim* is to draw upon certain of Bernard Williams' influential remarks in his articulation of the 'one-thought-too-many' argument.[19] The general suggestion, here, is that the singular thought criterion ensures that our practical deliberations about how to treat others can be *directly motivated* by a concern for those who stand to be affected by our choices.

Roughly speaking, the idea goes as follows. If I designate individuals using generic labels which only facilitate general thought – labels such as, say, *The Person I Will Kill By*

[18] Of course, the claim is not the obviously erroneous one that I somehow have the power to make it the case that there are separate rather than non-separate persons affected by my choice. The claim is simply that conforming to the singular thought criterion is a manner in which decision makers can *respond appropriately* to the separateness of persons in the course of their moral deliberations.

[19] Williams (1981: 1–19). What I articulate in these paragraphs is clearly not the one-thought-too-many argument itself, but rather, an argument which, in the same vein, emphasises the important role that directness of motivation might be thought to play in practical reason and the realm of interpersonal morality.

Pushing – then my practical deliberation about whether or not to *Push* is inevitably going to have to go through some abstract chain of reasoning which references general moral principles specifying how I ought to treat individuals in general. Perhaps my practical deliberation will have to go through an abstract chain of reasoning of the following sort: "Whenever I could do harm to this-or-that person, I ought not do so; this is a situation in which I could *ex hypothesi* do harm to this-or-that person; ergo, I ought not do harm to this-or-that person."

To reason in this way, one might worry, is to have *one thought too many*. It is for my practical reasoning to be grounded in abstract principles which stipulate how I ought to treat individuals in general. By contrast, if I designate individuals using labels which facilitate singular thought – such as *Alice* – then it seems that my deliberations can spring directly from concern for the *particular individual* about whom I find myself thinking, namely, Alice. I need not be compelled to action by some abstract principle which specifies how I ought to treat people in suitcases in general.

Those, then, are two different ways in which we might further buttress the *Concern Claim*. Singular thoughts allow the demonstration of a special and normatively important kind of concern. For they allow agents to deliberate *separately* about the individuals who stand to be affected by their choice. And they also allow agents to be *directly motivated*, in their practical deliberations, by the patients who stand to be affected by their choice. That is why, insofar as they are able to, agents should employ designators conducive to singular thought.[20]

It may be worth pausing here to emphasise what I have *not* claimed in the above discussion. I have not claimed, above, that it is impossible for me to show concern for an

[20] Perhaps these two rationales are closely connected, such that they stand or fall together: perhaps, for instance, being directly motivated by the plight of someone else standardly requires respecting the separateness of persons.

individual *i* insofar as I am incapable of thinking of that individual singularly. I can clearly show concern for *The Tallest Person In Future Spain*, even though, under standard analyses of singular thought, I am incapable of thinking about such a person singularly. Learning that such a person will be assassinated on high noon upon the turn of the next century, I might pay a great deal of money to buy out the assassin in question and thus prevent their deadly plot from coming to fruition. In doing so, I clearly show *The Tallest Person In Future Spain* concern. All of this does nothing to contradict the central suggestion I have considered here: namely, the suggestion that I ought to use designators which facilitate singular thought insofar as I can do so, since to do so is to facilitate a *normatively special* kind of concern towards other persons.[21]

3.4 Two Observations

Before concluding discussion of the singular thought criterion, let me make two important observations. The first observation goes as follows. In my defence of the singular thought criterion, we have arrived at an answer to what we might call the *Perception Question*, articulated forcefully in recent work by Kieran Setiya. In Setiya's recent analysis of *personal acquaintance*, where personal acquaintance is whatever cognitive relation *R* it is which justifies love (at first sight) and personal concern (at first sight), Setiya at certain points despairs at the normative importance of mere perceptual contact. In particular, Setiya asks: "Why should perceptual contact, past or present, have moral significance? Why should seeing the person on the bridge, or being able to identify them with someone

[21] N.B. Setiya (2023). In particular, Setiya (2023: 323) draws our attention to the fact that interpersonal concern for others with whom I am *personally acquainted* appears to have a special ethical dimension, but this is not to commit us to the further conclusion that this is the only type of concern attitude which might appropriately feature in the normative landscape: "The idea is rather that concern mediated by personal acquaintance has an ethical significance that is not shared by concern for the person, whoever it is, that meets a given description."

you met before, make it wrong to push the button, where knowing *that you will kill the person on the bridge* does not?"[22]

The concern raised here has a great deal of intuitive force. Cast in a certain light, the suggestion that seeing a patient (rather than merely knowing that such-and-such a patient exists) could bear upon one's deontic obligations seems bizarre; such a result seems powerful enough to serve as a *reductio* of any framework which entails it.

Nonetheless, we now have the equipment in place to develop an answer to the Perception Question. Perceptual contact with patients can indeed make a difference to the deontic status of your acts. This is not, however, because perceptual contact is of fundamental normative importance *per se*. It is because agents have normative reason to use those modes of designation (when allocating prospects) which are capable of facilitating a distinctive and normatively important kind of singular thought about patients, a kind of thought partly constitutive of *what it is* to treat patients as separate persons and to respond directly to their interests. It just so happens to be a contingent feature of our psychologies, given the sorts of creatures we are, that whether or not I can think singularly about an individual i (under a given mode of designation D) is at least partly sensitive to whether I know myself to have had perceptual contact with i (presented under such a mode of designation D).

One way of emphasising that it is not perceptual contact *per se* which is of fundamental normative importance is to note that, if it turned out that singular thought could come about in virtue of mechanisms of acquaintance *other* than perceptual contact – a possibility I previously left open – then singular thoughts emerging in virtue of these other mechanisms would nonetheless be morally desirable in the special sense I have indicated

[22] Setiya (2020: 71, emphasis. added).

too.[23] Perceptual contact would thus be sufficient (but not necessary) for facilitating the morally desirable thoughts in question.[24]

The second general observation to make is that, clearly, in my presentation of the singular thought criterion, I have skipped over certain difficult edge cases – edge cases in which it is unclear whether I have had perceptual contact sufficient for forming singular thoughts about patients.

Suppose, for instance, that I can only see some individual's left elbow. Do I have perceptual acquaintance with that individual sufficient for baptising them as *The Patient* and for thinking singularly of such a person? May I now allocate prospects to *The Patient* in a manner which would previously have been impermissible? What, too, about hazy reflections in disturbed puddles? Faint voices echoing across a crackly Ham radio?

It seems to me that we may be able to say something concrete and constructive, here, about how such edge cases ought to be handled. The relevant question to ask about such cases goes as follows: "Would catching a glimpse of *The Patient's* elbow suffice for me to be able to think about *The Patient* (so mentally presented) as a separate person, rather than a certain kind of 'mental disjunction' of persons? Would catching such a glimpse be sufficient for me to be motivated by a direct concern *for The Patient* (so mentally presented)?" If the answer to these questions is *yes*, then it is permissible to allocate prospects to *The Patient*, even if I have only (say) seen their left elbow.

I am inclined to think that the answer is to these questions is indeed yes: such edge cases of perceptual contact *are* sufficient for generating normatively important thoughts

[23] See, for instance, the *Standard Standard* view in fn15., which allows singular concepts to be transmitted through testimonial chains so long as those chains end in perceptual contact.
[24] Setiya's (2023) own view is that there does indeed exist a relation of 'personal acquaintance', *R*, such that the presence of *R* is what justifies love (at first sight) and personal concern (at first sight). This is even though it may be impossible to give some further analysis of *R* in terms other or more fundamental than these, and even if personal acquaintance is not expressible in terms of, say, perceptual acquaintance.

of the relevant kind. The explanation of this inclination may lie in the fact that, in such cases, even brief or distorted or imperfect perceptual contact with the patient allows me to structure my thought in a directed way which I could not do previously – in a way which is now about *that patient, the one over there.* It goes without saying, however, that one need not share these inclinations in order to accept that the singular thought criterion should govern the activity of naming when allocating prospects.

4. The Informational Criterion

That, then, is one criterion of *ex ante* designation. Insofar as it is possible, agents ought to designate the individuals affected by their choice by making use of designators which facilitate singular thought about those individuals.

Note, though, that the singular thought criterion alone will not be enough to completely solve the problem emerging in a case like the original **Suitcases**. To see why, consider the following point. In the original statement of **Suitcases**, I asked what you should do if you know that the person in a suitcase on the bridge commonly goes by the nickname *Ghost Rider*, where such a person enjoys the occasional game of football. We can suppose that you have become acquainted with Alice (in the sense relevant for singular thought about Alice) and acquainted with *Ghost Rider* (in the sense relevant for singular thought about *Ghost Rider*) without knowing whether Alice is in fact *Ghost Rider*. In such a case, which of these designators ought you use as the basis for the allocation of prospects? As before, different answers will change the deontic status of one's options under *ex ante* Pareto.[25]

[25] We should emphasise, however, that the singular thought criterion nevertheless winnows away very many gerrymandered designators – designators such as *The Person Who Will Die If I Push.* Countless such gerrymandered designators, presumably, would emerge in any case deterministic case.

In this section, I consider a second criterion which will be of use in settling these questions. I call this second criterion the *informational criterion*. A useful thing about the informational criterion is that its plausibility emerges reasonably independently of the singular thought criterion. Thus, those who find themselves unpersuaded by my first criterion may still avail themselves of the second.

4.1 The Informational Criterion

Let me begin by stating the informational criterion outright:

> **The Informational Criterion.** At the *ex ante* stage of moral decision making, agents have reason to use those designators which are (as things seem to the agent themselves) *maximally informationally loaded*.[26]

This statement of the criterion will require some finessing. For now, here is a general heuristic for discovering how informationally loaded a designator *D* is to a decision-making agent *A* at the *ex ante* stage. We simply ask the agent *A* to tell us everything they know, at the *ex ante* stage, to be determinately true of the individual picked out by the designator *D*.

Suppose, in the original **Suitcases**, you know a great deal about each of the individuals known as Alice, Blaise, Charlie … and so on. (You know, for instance, that Alice was born in Albania, emigrated to Albuquerque, enjoys art, works as an architect, has a pet ant … and so on). Suppose, by contrast, that you know nothing at all about the individual who goes by the moniker *Ghost Rider* (except that they enjoy the occasional game of football).

If this is the structure of the case, then the informational criterion speaks in favour of your allocating prospects to individuals designated by proper names (like Alice) rather

[26] A similar idea is footnoted and rejected in Mahtani (2021: 2086). As will become apparent, in light of my suggested rationale for the informational criterion, I don't think we should be worried by Mahtani's concern about the garnering of irrelevant information.

than designated by monikers such as *Ghost Rider*. For if we asked you to list everything you knew about the individual who goes by the name 'Alice' you would be able to tell us a great deal; by contrast, if we asked you to tell us everything you knew about the individual who goes by the moniker *Ghost Rider* you would barely be able to say anything at all.

That, then, is an initial sketch of the informational criterion and of how it might be applied to a case like **Suitcases**. Now, it turns out that the informational criterion is already in need of finessing. Suppose, for instance, that you knew the exact spatial location of each of the blood cells contained within *Ghost Rider's* left thumb. This is to know an *enormous* amount of information about the individual picked out by the label *Ghost Rider*; it is to know trillions of facts about the exact biological composition of such a person. But it does not seem that learning this sort of information should in any way tip the scales, making *Ghost Rider* a more worthy designator than (say) the designators Alice, Blaise, Charlie, and so on. It does not seem that learning this information should, in brief, bear upon the question of how you ought to designate or what you ought to do.[27]

How to resolve this puzzle? It ultimately seems that only a certain *kind* of information should count as relevant in determining the 'informational load' of a designator. Here is the particular hypothesis I have in mind. The particular kind of information which matters, in determining a designator's informational load, is information which somehow speaks to the relevant patient's *individual identity*.[28]

---

[27] *Ibid.*

[28] I borrow this label from Mogensen (2021), who, by contrast, provides an error theory about the cogency of the notion. The concept of individual identity is not to be confused with the ubiquitously employed concept of *personal identity*, where that latter concept attempts to pick out whatever relation it is (whether psychological, physical, or some other) which accounts for the continued survival of a given individual over time. As Mogensen (2021: 57), notes, various historical debates presuppose either the notion of individual identity or some sufficiently close analogue. See, for instance, discussions of authenticity in Taylor (1991) and discussions of blame in Sher (2005).

The concept of individual identity attempts to pick out the following idea. As a human being, I am individuated by countless properties – my hair colour, my coincidental distance from various celestial objects, my religious and political beliefs, my tendency to enjoy music rather than painting, and so on. It is extremely natural, however, to maintain that some of these properties play a more central role in making me *me*, or in accounting, to put the point another way, for the sort of person I am. My religious and political beliefs are in this exact sense more central to my individual identity than is my coincidental distance at any given point in time from various celestial objects. The current proposal is simply that, when measuring the 'informational load' of a designator, we should assign greater weight to those properties which speak to someone's individual identity. It is this fact which explains why it would nonetheless be inappropriate to use the designator *Ghost Rider* to pick out patients in **Suitcases**, even if I knew trillions of facts about (say) the exact spatial location of the blood cells within *Ghost Rider's* left thumb. Such facts simply do not speak, in any comprehensible way, to *Ghost Rider's* individual identity.

Note that, in saying that properties can be more or less central to our individual identity, we need not commit ourselves to *metaphysical essentialism* about identity; in particular, to the view that I would not been me had I lacked any of these properties. We need only commit to the idea that, of the various properties which I actually possess, some of these properties (such as deeply held political or religious beliefs) play a greater role than others (such as my distance from celestial objects) in accounting for the sort of person I am. One particular way of precisifying the idea, to which I am sympathetic, is that those properties more central to one's individual identity feature in the narrative account that each of us would give in explaining our life stories to either ourselves or others.[29] Such an

[29] See, for instance, Schechtman (1996).

account can clearly come apart from metaphysical essentialism, since persons can tell their own narratives without making claims that they have such a thing as a 'necessary essence.'

When it is precisified to be responsive to the notion of individual identity, the informational criterion reads as follows. An *ex ante* designator is more highly 'informationally loaded' insofar as, when I employ that designator, I know a greater quantity of identity-relevant information about the individual who is in fact picked out by that designator. This finessed statement of the criterion explains why, when ranking *ex ante* designators, designators loaded with enormous quantities of information about the distribution of blood cells in a given patient's left thumb score comparatively lowly, while designators loaded with information about someone's family, their work, their religious beliefs, their value systems, and so on, score comparatively highly. In normal conditions, these properties play a significant role in one's individual identity. Not so, for the trajectory of thumb-bound blood cells.

4.2 In Favour of the Informational Criterion

Why should agents designate such that they satisfy the informational criterion? As with the singular thought criterion, it seems to me that we can support the informational criterion with a principled rationale.

The general idea, in brief, is that we as patients have a certain kind of interest in being *recognised* by decision makers, an interest towards which decision makers ought to be responsive when making choices which affect us. Decision makers recognise patients insofar as they use modes of designation loaded with more (rather than less) identity-relevant information. So decision makers ought to use modes of designation loaded with more (rather than less) identity-relevant information. Call this argument, for want of a label, the *recognition-based argument* in favour of the informational criterion.

It is easiest to develop the recognition-based argument, and the concept of 'recognition' upon which it draws, by returning to the original **Suitcases.** Suppose that I am in fact patient Alice. I am entirely clueless as to which particular suitcase I am locked inside. The agent who holds my life in the balance is entirely clueless as to which particular suitcase I am locked inside. But the agent does know, as we have granted, a great deal of information about the individual identity of the person who goes by the name "Alice": the agent could tell you, if asked, about Alice's country of origin, about Alice's profession, about Alice's interests, about Alice's hobbies, and even about Alice's taste in pets. (*Mutatis mutandis* for Blaise, Charlie … and so on).

Imagine I learn, just as the agent is about to make their choice, that the agent *did not*, in reasoning about what to do, calculate prospects by making use of informationally rich designators such as "Alice". Instead, the agent deliberated about what to do by calculating prospects for *Ghost Rider*, knowing nothing more about *Ghost Rider* than that such a person enjoys the occasional game of football.

It seems to me that, in such a situation, I would have a certain kind of complaint against the agent. What, exactly, does this complaint amount to? Well, here is a truism: as a patient whose welfare is at stake, I want the agent to take account of my welfare in those decisions which stand to affect me. But more than this, it seems that I want the agent to take account of *me in particular* in their moral decision making. I want the agent to bear my rich individual identity in mind, insofar as it is reasonably possible, when making those decisions which stand to affect my welfare. Such an interest is satisfied insofar as the agent uses more (rather than less) informationally-loaded designators. This is what I mean, in

short, when I say that patients have an interest in being *recognised* by those decision makers who hold sway over their welfare.[30]

The interest of recognition is closely related to (but distinct from) our interest in being respected. Under a fairly intuitive gloss of that word, I am respected as a patient insofar as I am treated as an end rather than as a mere means to an end. We might say that the satisfaction of our interest in being recognised requires more than simply being respected: I am recognised as a patient insofar as I am responded to as an end with a distinctive identity, a distinctive identity which ought to be held in mind when others make decisions which stand to affect me.

Do we really care that patients be 'recognised' in such a way? It seems to me plausible that we care about recognition (of this particular kind) intrinsically. But even if the value of recognition were shown to be non-intrinsic, recognition is nonetheless crucial for securing certain *other* moral goods which are routinely taken to be of intrinsic value. So I would still maintain that agents should conform to the informational criterion in *ex ante* moral decision making.

Here is one example. Many of us care, on some fundamental level, that patients are treated as autonomous and self-determining beings. But insofar as we care about treating patients as autonomous and self-determining beings, we must do our best to *recognise* those patients in the sense I have outlined above. Why? Plausibly, what it means to treat some patient *p* as an autonomous and self-determining being is in part to be responsive to the sorts of preferences *p* actually has or would reasonably hypothetically form, such that,

[30] Importantly, the claim is clearly not that I would have an all-things-considered *ex post* complaint if they pushed the suitcase in a manner which ultimately saved my life. The claim is simply that, prior to the discovery of choice or outcomes, there is a distinctive kind of *ex ante* complaint I can level against another if she has failed to make her decision while bearing in mind what makes me *me* in her choice. This broad distinction about the types of complaints one can have is ubiquitous, and so it shouldn't be problematic if we make use of it here: I might have an *ex ante* complaint against a decision maker who implemented a policy which lowered my chances of survival but an *ex post* 'gratitude' that the policy in-fact ended up saving my life.

in my treatment of them, I do not do what they actively disprefer me to do. But it is hard to believe that this kind of respect for autonomy or self-determination will standardly be possible insofar as I deliberately choose those modes of mental presentation, for *p*, under which her individual identity is 'screened off.' It is not possible, as a practical matter, to be responsive to the preferences and natures of individuals whom I have chosen to mentally present as a mere *Person In A Suitcase*. Thus, even if one were sceptical that recognition is of intrinsic moral importance, something akin to recognition does seem necessary in order to secure other valuable aspects of interpersonal morality, such as the interests of autonomy and self-determination.[31]

In sum, then: I am robustly recognised insofar as the agent uses the informationally-rich label "Alice." I am only shallowly recognised insofar as the agent uses the informationally-thin label *Ghost Rider*. That is why the agent in question has a normative consideration in favour of calculating prospects for "Alice" in deliberating whether or not to *Push*. This would not have been the case if the agent knew no details at all about the individuals picked out by the proper names Alice, Blaise, Charlie, and so on. If that were the structure of the case, then that normative consideration in favour of using those names would largely disappear.

5. Criterial Conflict

That, then, is a theory of *ex ante* designation. Agents ought to use those designators capable of featuring in singular thoughts. And agents ought to use those designators which are 'maximally loaded' with information relevant to the individual identity of patients. For each such criterion, I've shown how that criterion can be underpinned by a normative rationale.

[31] For related themes, see Raz (1988).

In principle, these two criteria could conflict. One set of designators $D_1 - D_n$ might facilitate singular thought about potential victims; another set of designators $E_1 - E_n$ might be especially loaded with identity-relevant information about victims. In such cases, what ought one do? Which mode of designation ought one employ?

My own view is that these are cases in which there may be no determinate answer as to how one ought to designate the individuals affected by one's choice, and subsequently, as to whether or not one's action is permissible when evaluated from the *ex ante* standpoint. This is, I think, an upshot we should accept. For it ultimately appears that these are cases in which different modes of designation facilitate different facets of interpersonal morality which we each take to be of value. One possibility, for instance, is that the different modes of designation within such cases stand in a parity relation with one another: insofar as each mode of designation facilitates a different valuable facet of interpersonal morality, neither mode of designation is better nor worse nor exactly equally as desirable as the other.[32]

One might find this suggested conclusion unconvincing. In particular: doesn't this mean that we've ended up back where we started? After all, we began with the problem that there are certain cases in which it's unclear how to designate patients. And we've now *ended* with the problem that there are certain cases in which it's unclear how to designate patients.

This complaint isn't quite right, however. We clearly *have* arrived at a very different position from the one in which we started. We have specified what are the norms which ought to govern the activity of naming when allocating prospects. And we have

[32] Famously, see Chang (2002). In deriving prescriptions for such cases one might, following Mahtani (2017), employ a supervaluationist rule such that an act is *ex ante* Pareto superior so long as it satisfies *ex ante* Pareto under each candidate mode of designation.

discovered, upon reflection, that there are certain cases with a peculiar structure (designators $D_1 - D_n$ score highly on the singular thought criterion and designators $E_1 - E_n$ score highly on the informational criterion) such that the relevant normative criteria conflict. There is a clear difference between recognising ourselves to be in this position, and between finding ourselves simply clueless about how to proceed in the activity of designation *simpliciter*, as we were at the start.

This complaint is especially undermined, I think, when we turn to observe that, for a very large range of practically relevant cases, the singular thought criterion and the informational criterion end up speaking in unison. In the event that I *know* Alice, in virtue of the two of us having some sort of shared history, then both the singular thought criterion and the informational criterion will speak in favour of designating using the label *Alice* (rather than designators like *The Person In A Suitcase On The Bridge*). Using the label *Alice* will let me think of that person singularly in making my choice. And since I know much about the individual identity of the one called *Alice*, this label will satisfy the informational criterion too. This, then, is the label I ought to use.

6. Concluding Remarks

The theory of *ex ante* designation I've sketched here is an optimistic one. It says that there is often a uniquely ethically acceptable way of designating patients when allocating prospects. Of course, as we've seen, the proposed theory sometimes has surprising implications. One surprising implication was that perceptual contact with patients can make a difference to how I ought to allocate prospects. After all: if I have become perceptually acquainted with the person who goes by the name *Ghost Rider*, this might open up the possibility of singular thought about *Ghost Rider*, and I have reason to use those designators which allow me to think about persons singularly. Another surprising

implication was that learning a long list of identity-relevant facts about *Ghost Rider* (their political beliefs, their family background, and so on) could make a difference to how I ought to allocate prospects. After all: I have reason to use those designators which are highly 'loaded' with identity-relevant facts about the individuals affected by my choice, identity-relevant facts just such as these.

Some may find these practical consequences simply unacceptable. On the one hand, certain *ex ante* theorists might hold out hope for a different theory of *ex ante* designation, a theory of designation which would avoid entailing these practical consequences. On the other hand, *ex post* theorists might be tempted to take these practical consequences as a *reductio* against their *ex ante* opponents. I tentatively fall into neither camp – I am tentatively willing to entertain both the theory and its consequences.

References

Adler, M. (2011) *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis*. Oxford University Press.

Armstrong, J. & & Stanley, J. (2011) 'Singular Thoughts and Singular Propositions', *Philosophical Studies*, 154: 205–222.

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

Crane, T. (2011) 'The Singularity of Singular Thought', *Proceedings of the Aristotelean Society Supplementary Volume*, 85: 21–43.

Fischer, J. (2024) 'Consequentialism and the Separateness of Persons', *The Philosophical Quarterly*, 1–24.

Frick, J. (2015) 'Contractualism and Social Risk', *Philosophy & Public Affairs*, 43(3): 175–223.

Glover, J. (1977) *Causing Death and Saving Lives*. Penguin.

Gustafsson, J. E., & Kowalczyk, K. (forthcoming) 'Ex-Ante Pareto and the Opaque-Identity Puzzle', *The Journal of Philosophy*, 1–15.

Hare, C. (2016) 'Should We Wish Well to All?', *Philosophical Review*, 125(4): 451–472.

Harris, J. (1975) 'The Survival Lottery', *Philosophy*, 50(1): 81–87.

Horton, J. (2017) 'Aggregation, Complaints, and Risk', *Philosophy & Public Affairs*, 45(1): 54–81.

Jeshion, R. (2010) 'Singular Thought: Acquaintance, Semantic Instrumentalism, and Cognitivism', in R. Jeshion (ed), *New Essays on Singular Thought*. Oxford University Press.

Kamm, F. (1996) *Morality, Mortality: Volume II: Rights, Duties, and Status*. Oxford University Press.

Kaplan, D. (1989) 'Demonstratives', in J. Almog, H. Wettstein, and J. Perry (eds), *Themes from Kaplan*. Oxford University Press.

Mahtani, A. (2017) 'The Ex Ante Pareto Principle', *The Journal of Philosophy*, 114(6): 303–323.

——— (2021) 'Frege's Puzzle and the Ex Ante Pareto Principle', *Philosophical Studies*, 178(6): 2077–2100.

McCarthy, D. (2006) 'Utilitarianism and Prioritarianism I', *Economics and Philosophy*, 22(3): 335–363.

——— (2008) 'Utilitarianism and Prioritarianism II', *Economics and Philosophy*, 24(1): 1–33.

Mogensen, A. (2021) 'Is Identity Illusory?', *European Journal of Philosophy*, 29(1): 55–73.

Nagel, T. (1970) *The Possibility of Altruism*. Princeton University Press.

Nebel, J. (2020) 'Rank-Weighted Utilitarianism and the Veil of Ignorance', *Ethics*, 131(1): 87–106.

Nozick, R. (1974) *Anarchy, State, and Utopia*. Basic Books.

Oberdiek, J. (2017) *Imposing Risk: A Normative Framework*. Oxford University Press.

Rawls, J. (1971) *A Theory of Justice*. Harvard University Press.

Raz, J. (1988) *The Morality of Freedom*. Oxford University Press.

Rüger, K. (2018) "On Ex Ante Contractualism," *Journal of Ethics and Social Philosophy*, 13(3): 240–258.

Schechtman, M. (1996) *The Constitution of Selves*. Cornell University Press.

Setiya, K. (2020) 'Ignorance, Beneficence, and Rights', *Journal of Moral Philosophy*, 17(1): 56–74.

——— (2023) 'Other People', in S. Buss & N. Theunissen (eds), *Rethinking the Value of Humanity*. Oxford University Press.
Sher, G. (2005) *In Praise of Blame*. Oxford University Press.
Taylor, C. (1991) *The Ethics of Authenticity*. Harvard University Press.
Temkin, L. (2009) 'Illuminating Egalitarianism', in T. Christiano & J. Christman (eds), *Contemporary Debates in Political Philosophy*. Wiley-Blackwell.
Voorhoeve, A. & Fleurbaey, M. (2013) 'Decide as You Would With Full Information! An Argument Against *Ex Ante* Pareto', in *N. Eyal, S. Hurst, O. Norheim, and D. Wikler* (eds), *Inequalities in Health: Concepts, Measures, and Ethics*. Oxford University Press.
Walen, A. (2020) 'Risks and Weak Aggregation: Why Different Models of Risk Suit Different Types of Cases', *Ethics*, 131(1): 62–86.
Williams, B. (1981) 'Persons, Character, and Morality', in J. Rachels (ed), *Moral Luck*. Cambridge University Press.

7.

Concluding Remarks

The essays in this collection were diverse. Nevertheless, they were unified by some close thematic connections. Broadly speaking, I began by asking about the *limits* of beneficence: that is, by asking about the extent to which each of us must structure our lives around charitable giving and behaviour. I then turned to ask about the *nature* of beneficence: that is, about some of the ways in which we should (or should not) engage in charitable giving and behaviour.

Chapter 1 developed a novel and revisionist theory of beneficence, a theory according to which each of us must begin our adult moral lives by devoting ourselves towards *altruistically-focussed lifeplans*. I showed how this central result followed from ecumenical premises, and how it offered to illuminate various existing debates about the limits of beneficence – most notably, the familiar debate over the extent to which we must each sacrifice our disposable income towards charity.

Although the reasoning within Chapter 1 was ecumenical, as the essays went on, I often felt the need to adopt an explicitly non-consequentialist perspective on how beneficence ought to work. This was most apparent in Chapters 3 and 5. In Chapter 3, for instance, I argued that altruistic near-bias is generally permissible across the lives of beneficiaries but generally impermissible within the lives of beneficiaries. In suggesting that considerations of partiality can indeed ground near-biased patterns of concern across future generations, I made a move which few impartially-minded consequentialists were likely to find appealing. And in Chapter 5, in my defence of *ex ante* deontology, I explicitly endorsed the idea that our attempts to promote the good are governed by deontic

constraints: I suggested that, sometimes, we are morally prohibited from bringing about those outcomes which would be impartially best.

Of course, the previous chapters haven't been a thoroughgoing attempt to vindicate all aspects of contemporary non-consequentialist thought. In Chapter 2, I ultimately became sceptical of the idea that we can have autonomy-based moral reasons for performing acts which facilitate our access towards distinctively valuable prudential features. I became sceptical of this idea, and of the moral autonomy movement more generally, insofar as these claims seemed to systematically entail lesser number permissions, an implication which I found very difficult to accept. And in Chapter 4, my aim was neither to prove nor disprove popular non-consequentialist sentiments about how we should partially aggregate claims for aid: my aim, rather, was to figure out what partially aggregative principles would imply when extended to cover the practice of charitable giving.

One theme which did become clear, towards the end, was a general sympathy towards *ex ante* principles as guides towards beneficence and non-maleficence. As things seemed to me, *ex ante* principles weren't just highly intuitively appealing. They also withstood some recently influential objections, such as the diachronic objection to *ex ante* deontology explored in Chapter 5 and the puzzle of *ex ante* designation explored in Chapter 6. In this way, the prospects for *ex ante* principles ended up looking quite good.

It's easy to sit in a comfortable chair, drinking coffee and typing away about how to do good in the world. One hopes that philosophical contributions such as these, though, are small steps forward in actually doing that good, and in doing it in the right way.