

Consciousness and the Good



Mattia Cecchinato

St Cross College
University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

30th July 2026

Abstract

This thesis is a collection of articles on the moral and prudential significance of consciousness. It investigates the ways in which the very fact that an individual has a conscious mental life determines (i) whether anything can be good or bad for them, (ii) what, if anything, is good or bad for them, (iii) whether they have prudential or self-interested reasons to care about their own lives or the lives of others, and (iv) whether they matter morally in and of themselves. It also discusses how these facts relate to our reasons to care about future generations, animals, the dead, and artificial intelligence systems. Chapter 1 argues that introspection gives us strong reason to believe that only conscious beings have lives that can go well or badly for them. Chapter 2 argues that consciousness itself, while necessary, is not sufficient for having a life that can go well or badly, and thus for mattering morally. What is required is a particular kind of consciousness, the affective kind possessed by beings capable of pleasure, pain, and emotion. The chapter argues that this has important implications for the moral status of animals and the future design of artificial intelligence systems. Chapter 3 argues that even if only conscious beings matter morally, there is still an important sense in which beings who were once conscious can be harmed or benefited after death. This is because something of their mental life can survive in other minds, namely the psychological connections that grounded the prudential concern they once had for themselves. Chapter 4 argues that because of these surviving psychological connections, we should care about the wellbeing of future generations for the same sorts of prudential reasons we care about our own. This means not only that our deaths may be less bad than we thought, but also that the extinction of humanity could be worse.

Word Count: 62,010

Acknowledgements

For at least a decade now, I have been trying to learn how to do philosophy a little better, and to get closer to what I hope is the truth about certain philosophical questions that seem to me of startling, almost paralysing, importance. Philosophy for me is not simply a collection of fun puzzles. It seems to me that the value of my own life, and the value of the entire world—of everything that exists, and everything that could exist—depend on the answers to such questions. Given what these questions mean to me, reflecting on them in a place like Oxford has been a tremendous gift. It is one of the things that, on my deathbed, I imagine I will be grateful to have lived. For even though I still feel tremendously far from finding the answers I am looking for, and even though at times, during my doctorate, the answers have seemed more distant than ever, almost beyond reach, there have also been times when they have never felt so close. So clearly within reach.

Those to whom I owe the greatest thanks for this are my DPhil supervisors, Roger Crisp, Jeff McMahan, and Andreas Mogensen. With inexhaustible patience, they have helped me develop my ideas in their strongest form, pointed out the toughest challenges to each of them, puzzled me when I was unpuzzled, and unpuzzled me when I was puzzled. They have taught me not only when it is time to be intellectually humble, but also when it is time to be intellectually brave. Above all, they have given me a sense of the kind of philosopher I would like to become myself one day.

A special thanks also goes to my former MPhil supervisor, Theron Pummer, whose incredibly sharp feedback greatly shaped the early versions of two chapters of this thesis—Chapters 3 and 4—which were originally written for my MPhil at St Andrews.

There are also a few communities and institutions to which I owe a great debt. I am especially indebted to the Uehiro Oxford Institute, the global

priorities research community, and the ethics of consciousness community, all of which have deeply influenced the ideas in this thesis and, perhaps more importantly, my view of what moral philosophy can and should do. I am grateful to Jesus College and St Cross College for having been wonderful places to work and live throughout the doctorate. I have also been fortunate to be supported by an Open Philanthropy Scholarship, a St Cross Uehiro Scholarship for Future Generations, and an Oxford Global Priorities Fellowship.

Along the way, I have had the privilege of thinking through the questions at the heart of this thesis with many friends and colleagues. For especially precious conversations, I am grateful to Uriah Kriegel, Neil Levy, Andrew Lee, Willem van der Deijl, Gwen Bradford, Julia Driver, David Brink, Dan Pallies, Jeff Sebo, Joshua Shepherd, Jonas Aaron, Guy Kahane, Richard Yetter Chappell, Helen Yetter-Chappell, Eden Lin, Tomi Francis, Louise Mackie, Joel Joseph, Filippo Ferrari, Tom Kaspers, James Glover, Brad Saad, Patrick Butlin, Luca Stroppa, Bob Fischer, Ewan White, Jen Semler, Melania Pellitteri, Hayden Wilkinson, Tim Mulgan, Kida Lin, Paul Heller, Lorenzo Elijah, Tom Douglas, Riccardo Valsecchi, Dario Dedi, Patrick Williamson, Rhys Southan, Rei Takahashi, Ryan Kulesa, Lorenza D'Angelo, Gary O'Brien, Calvin Baker, Lucius Caviola, Theodore Naylor, Elad Uzan, Riley Harris, Pietro Cibinel, Harry Lloyd, Jakob Lohmar, and Elisabetta Sassarini. There will be someone I have forgotten. I hope they'll understand. It's been just such a rich period of my life.

Table of Contents

Abstract	2
Acknowledgements	3
Table of Contents	5
Introduction	7
1. <i>No Consciousness, No Welfare</i>	12
1.1 Introduction	13
1.2 Introspection and Justification	18
1.3 The Argument from the Night (Pt. 1).....	20
1.3.1 Losing Consciousness.....	21
1.3.2 Regaining Consciousness	28
1.4 The Argument from the Night (Pt. 2).....	30
1.5 The Argument from the Night (Pt. 3).....	35
1.6 Defending Night	39
1.6.1 Undercutting Defeaters.....	40
1.6.2 Rebutting Defeaters	44
1.7 Conclusion	54
2. <i>The Mind that Matters</i>	55
2.1 Introduction	56
2.2 The Argument for Affective Sentientism.....	60
2.3 Vulcans, Desires, and Caring	67
2.4 Other Goods	75
2.5 Objections.....	82
2.5.1 Vulcanization and Death.....	82
2.5.2 Vulcanization and Buddhism.....	85
2.5.3 Killing Vulcans.....	87
2.5.4 Autonomy and Caring	88

2.5.5 Neutral Welfare and Welfare Subjects	89
2.6 Conclusion and Implications.....	94
3. <i>On What Matters to the Dead</i>	97
3.1 Introduction	98
3.2 Fission-Cases and Posthumous Harms.....	103
3.3 The Continuity View of Posthumous Harm.....	111
3.3.1 Interpersonal Psychological Continuities.....	111
3.3.2 Posthumous Psychological Continuities	117
3.3.3 Six Special Features	123
3.4 Objections and Further Problems.....	126
3.5 A Comparison with Other Theories	133
3.5.1 Psychological Continuity Plus Desire-Fulfilment.....	135
3.5.2 Psychological Continuity Without Desire-Fulfilment	138
3.5.3 Desire-Fulfilment Without Psychological Continuity	139
3.6 Conclusion	141
4. <i>Quasi-Surviving through Future Generations</i>	142
4.1 Introduction	143
4.2 Quasi-Survival.....	144
4.3 Two Strategies for Quasi-Survival	148
4.4 The Continuity Strategy	152
4.5 The Merits of the Continuity Strategy	159
4.6 Problems for the Continuity Strategy	166
4.6.1 The Dissimilarity Problem	166
4.6.2 The Decay Problem	169
4.6.3 The Disintegration Objection.....	172
4.7 Conclusion	174
Concluding Remarks	176
References.....	180

Introduction

That we have minds—or minds of a certain kind—is a fact of the utmost importance. It seems to be what makes us beings who matter morally for our own sake; beings who can be wronged, and who thereby have *moral status*. It may well be what makes us beings whose lives can go well or badly, who can be benefited or harmed, who have *wellbeing* or *welfare*. And the persistence of our mental life over time might be the ground of the special, prudential concern we typically have for ourselves. These are perhaps the most fundamental ways in which it matters morally and prudentially to have a mind, and the ones I try to understand in this thesis.

The central claim of the thesis is that what grounds a being's moral status and capacity for welfare is a certain kind of *phenomenally conscious* mind. Here, and in all that follows, phenomenal consciousness should be understood as subjective experience, such as the experience of seeing the redness of a rose or feeling the pleasantness of a warm sunny day.¹ Only beings capable of this kind of consciousness, I argue, have lives that can go well or badly for them, and so have interests that deserve moral consideration in and of themselves.

The thesis also investigates the grounds of prudential reasons, the kind of reasons we take ourselves to have to care about our own good, or our own welfare. I argue that the persistence of certain links between mental states is what gives us reason to care about a given individual in the distinctively egoistic, prudential way in which we typically care about ourselves. I also argue that, surprisingly, this gives us prudential reasons to care about the far future of humanity, and about many other things that will happen after we die. So even if the only bearers of welfare are conscious beings, our prudential reasons

¹ I take definition by example to be the best way of introducing phenomenal consciousness, since it fixes the relevant concept in a relatively theory-neutral way. For a defence of this approach, see Schwitzgebel (2016).

can extend beyond the boundaries of our own conscious life to include the conscious lives of others.

A couple of preliminary points are in order. First, I will not commit to any particular metaphysical view of the nature of conscious experience and its place in the physical universe. Rather, this thesis is about the moral and prudential significance of consciousness *whatever consciousness is*. This is because I think we have a much better grasp of the notion of conscious experience than of its metaphysics. My arguments are therefore meant to apply regardless of one's views about the nature of consciousness, though their implications may vary depending on what consciousness turns out to be. Second, while the thesis focuses on fundamental connections between certain facts about mentality and certain facts about moral and prudential value, it also discusses how these facts relate to our reasons to care about future generations, animals, the dead, and artificial intelligence systems. In this sense, the whole thesis is intended to have an applied dimension, with implications for a wide range of issues in practical ethics.

Let me say something about the structure of the thesis. This is an *integrated thesis*: a collection of four relatively independent but thematically interrelated chapters, each written in the format of a journal article and readable on its own. In what follows, I briefly explain the core argument of each chapter and how the chapters relate to one another.

In Chapter 1, *No Consciousness, No Welfare*, I present an epistemic argument for the view that consciousness is necessary for wellbeing. As we'll see, many believe, both inside and outside philosophy, that the capacity for consciousness is required for being the sort of creature that can be harmed or benefited, and that this is why rivers and large language models do not matter morally in the way that humans and other animals do. Yet there is no consensus on how exactly consciousness grounds welfare and moral status. And especially with the rise of artificial intelligence, some have begun to argue that sophisticated AI systems might have genuine welfare interests even without

being capable of consciousness.² In this chapter, I argue that we can justify the consciousness requirement on the basis of introspection. Specifically, I claim that the experience of gradually losing consciousness gives us reason to believe that only conscious beings can have wellbeing. I also argue that this introspectively justified belief is much harder to defeat than one might think. All things considered, then, we should accept it.

But even if consciousness is necessary for a life to matter morally, this doesn't settle the further question of whether it is sufficient. In Chapter 2, *The Mind that Matters*, I ask what particular kind of consciousness grounds wellbeing and moral status.³ Is it phenomenal consciousness per se, conscious agency, affective consciousness (pleasure, pain, and emotion), or something else? In existing humans and other animals, these features tend to go together. But in the near future, AI systems and brain organoids could develop minds vastly different from our own—minds that possess phenomenal consciousness and sophisticated agency without affective feelings. I argue that a being with such a mental life would be incapable of friendship, love, aesthetic appreciation, virtue, achievement, knowledge, and desire-fulfilment, at least of the kinds that contribute to wellbeing. Affective consciousness is the mind that matters, because every plausible constituent of wellbeing requires an affective component: an experience that feels good or bad, or a disposition to have such experiences.

Together, Chapters 1 and 2 can be taken as an extended defence of *Sentientism*, the popular view that all and only sentient beings have moral status.⁴ Chapter 2 proposes a qualified version, which I call *Affective Sentientism*, the view that all and only affectively conscious beings have moral

² On whether existing AI systems can have welfare even if they are not conscious, see Goldstein and Kirk-Giannini (2025).

³ This chapter is an extended version of my paper *The Mind that Matters* (forthcoming).

⁴ Explicit supporters include Singer (1977, 1993), DeGrazia (1996), Sumner (1996), Siewert (1998), Cutter (2017), Korsgaard (2018), Shepherd (2018), Kriegel (2019, forthcoming), van der Deijl (2021, 2025), Lin (2021), Lee (2022, 2025), Chalmers (2022), Nussbaum (2022), Roelofs (2023), Perry (2024), Fanciullo (2025b), Smithies (2026).

status, and defends it against alternative understandings of the sentientist commitment. It also argues at length that all and only beings capable of welfare have moral status, a connection that has recently been challenged.⁵ Both chapters, moreover, defend Sentientism without presupposing any particular theory of welfare. Neither is committed to *experientialism* about welfare (the view that all differences in wellbeing levels supervene on differences in conscious experiences), nor to *hedonism* (the view that pleasure and suffering are the only contributors to wellbeing). Chapter 2 concludes by investigating some of the practical implications of Affective Sentientism for the moral status of animals and the future design of artificial intelligence systems.

Beyond making us capable of wellbeing, the persistence of certain mental properties may also be what gives us reasons to care about that wellbeing in the distinctively egoistic way we do. Indeed, Derek Parfit in *Reasons and Persons* (1984) famously argued that the special prudential concern we typically have for our own good is grounded not in personal identity, but in certain psychological connections and continuities, such as continuities in memories, intentions, and beliefs. In the second part of the thesis, I examine the ethical implications of this view. I'm especially interested in one of its most revisionary features: that psychological connections and continuities can be easily established across different individuals. According to Parfit, a mental state A is connected to a mental state B if and only if there is (i) psychological similarity between A and B and (ii) a causal link of any kind responsible for that similarity. These conditions are often met in ordinary life: with relatives, partners, friends, and colleagues, we engage in ongoing causal interactions through shared experiences and conversations, cultivating deep similarities in interests, memories, tastes, and values. For prudential purposes, we are therefore connected to others in ways analogous to those by which we are connected to our past or future self. I argue that this feature of the Parfitian view is more plausible than has been recognised, and that it already explains

⁵ Kagan (2019), Mogensen (2025b).

our special attitudes towards future generations, the dead, and the threat of human extinction.

In Chapter 3, *On What Matters to the Dead*, I argue that the Parfitian view offers a solution to the long-standing puzzle of whether and how posthumous harm is possible (that is, harm occurring after death). Even if we can no longer harm individuals identical to the dead person, we can still harm others who are suitably psychologically related to her: the dead person's *continuers*. I develop this unconventional theory of posthumous harm and argue that it accounts for many of the familiar ways in which dead people are thought to suffer harm, while also opening up new ways to benefit them.

In Chapter 4, *Quasi-Surviving through Future Generations*, I ask whether these interpersonal psychological relations can hold between us and far future generations. I argue that they can: memories, intentions, beliefs, and values can be passed from one generation to the next, linking the lives of people who never coexist, and giving us prudential grounds to worry about the wellbeing of people who will exist only in the far future. This explains our intuitive responses to a number of puzzles about the badness of death and human extinction, and entails the surprising conclusion that we should care about the wellbeing of future generations for the same sorts of reasons we care about our own. Moreover, if future generations can carry on our mental life after we die, it turns out that our own deaths may be less bad than we thought.

Overall, the thesis argues that a certain kind of consciousness is what makes an entity the kind of being whose life can go well or badly for it. It gives a series of arguments to justify, motivate, and specify the best version of this widespread belief. But it also argues that certain connections between mental states can give us prudential, egoistic reason to care about lives that are, strictly speaking, not our own. To that extent, it departs from how we ordinarily think about self-interest.

Chapter 1:

No Consciousness, No Welfare

Both within and outside academic philosophy, it is widely believed that only conscious beings can be welfare subjects, subjects whose lives can go well or badly for them. Recently, however, this view—Phenomenal Necessitarianism—has come under sustained attack. In this chapter, I defend Phenomenal Necessitarianism with a new epistemic argument. My argument proceeds in two steps. First, I argue that the introspection of certain experiences seems to provide *prima facie*, defeasible justification for the belief that things can go well or badly for a subject at a given time only if it is conscious at that time. Phenomenal Necessitarianism follows from this. Second, I contend that there is no defeating evidence against this introspectively justified belief: neither from apparently opposing intuitions, nor from wellbeing theory, nor from philosophy of mind. Since in the absence of defeating evidence *prima facie* justification should be taken as all-things-considered justification, I conclude that we ought to be necessitarians.

1. Introduction

It's the year 2100. I'm walking down the street on a tranquil afternoon, when suddenly I hear desperate screams from just around the corner. I rush over. Around the corner is one of those Type-Y AIs, screaming as a kid hits its robotic legs with a stick. I pause, my heart racing. Now I remember. Type-Y AIs are not *phenomenally conscious* models. I am relieved. Type-Y AIs feel neither pain in their legs nor the warmth of sunlight on their cheeks. There is nothing it is like to be them. They experience nothing from within. Thus—I think to myself—they can be treated as we please without worrying about how they are doing or how things are going for them. The discovery that the victim is not a conscious being makes all the difference to me because I, like many others, believe

Phenomenal Necessitarianism. Being a conscious entity is necessary for being a welfare subject.

A welfare subject is a being whose life can go well or badly *for* it. It is a suitable bearer of *welfare goods or bads*—a being for whom certain things can be intrinsically beneficial or harmful, valuable or disvaluable, good or bad. Phenomenal Necessitarianism, then, is the view that *only conscious beings can be welfare subjects*. The necessity here should be understood as metaphysical necessity, and a conscious being as any being that has *phenomenally* conscious states or at least the capacity for them.⁶

Sometimes the difference between these two uses of 'conscious being' is important, so important that we should distinguish between:

State Necessitarianism. Being in a phenomenally conscious state at time T is necessary for being a welfare subject at T.

⁶ I'll remain neutral on what it is to have a capacity. With minor revisions, my arguments should be compatible with the most common analyses of the notion.

Capacity Necessitarianism. Being capable of phenomenal consciousness at time T is necessary for being a welfare subject at T.

These views yield different verdicts in cases involving comatose, sleeping, and fully conscious beings. Capacity Necessitarianism is the most widely endorsed, and the one I'll assume going forward, though I'll later suggest that a case can be made for State Necessitarianism.⁷ In either form, the necessitarian requirement has potentially far-reaching implications for public policy and pressing legal issues. Suppose—as is often claimed—that only welfare subjects have interests that matter morally in and of themselves, and hence *moral status*.⁸ Phenomenal Necessitarianism can then offer a bridge from results in the science of consciousness to the conclusion that certain AI systems, brain organoids, animals, or plants lack—or are likely to lack—interests that would otherwise demand some form of legal protection.

The necessitarian requirement is often combined with

Phenomenal Sufficientism. Being a conscious entity is sufficient for being a welfare subject.

The conjunction is known as *Sentientism*, the view that all and only conscious beings are welfare subjects.⁹ But the two commitments need not go together, and each can be evaluated on its own merits.¹⁰ In this chapter, I focus on the necessity claim as such: Why should we accept Phenomenal Necessitarianism

⁷ Explicit supporters include Singer (1977, 1993), DeGrazia (1996), Sumner (1996), Siewert (1998), Cutter (2017), Korsgaard (2018), Shepherd (2018), Kriegel (2019, forthcoming), van der Deijl (2021, 2025), Lin (2021), Lee (2022, 2025), Chalmers (2022), Nussbaum (2022), Roelofs (2023), Perry (2024), Fanciullo (2025b), Smithies (2026).

⁸ For a defence of this view, see Chapter 2; for an opposing view, see Mogensen (2025b).

⁹ This formulation of Sentientism can be found in van der Deijl (2021, 2025) and Lee (2022, 2025). But theorists who think that welfare subjectivity is grounded in a particular kind of consciousness, rather than in consciousness as such, specify Sentientism in other ways. On Affective Sentientism, for instance, all and only beings capable of *affective consciousness*—that is, consciousness that feels good or bad—are welfare subjects. See Chapter 2, and Smithies (2026).

¹⁰ Perhaps consciousness is a sufficient ground for welfare subjectivity but not the only one. Or perhaps consciousness is indeed required for welfare subjectivity, but bare phenomenology, without hedonic or motivational qualities, is not enough (Roelofs 2023; Smithies 2026).

(henceforth simply ‘Necessitarianism’)? What reasons do we have to believe that *only* conscious beings have welfare interests?

Three routes to justifying this requirement naturally come to mind. The first appeals to *intuition*—the mental state we are in when a verdict ‘strikes us’ as true.¹¹ The argument is simple: as long as the necessitarian requirement seems intuitively appealing, and intuitions provide justification for their content, the requirement is thereby justified. Some have indeed claimed to intuit that a creature behaviourally like us but entirely lacking conscious experience could not have a life that goes well or poorly for it.¹² Others, however, claim to intuit that sophisticated non-conscious AIs would count morally,¹³ or find it intuitive to speak of plant welfare without taking plants to be conscious.¹⁴ Given the extent of this disagreement, it seems that we can’t simply appeal to first-order intuitions; we also need to investigate the higher-order question of which intuitions are better justified. We need to examine how basic they are, whether their justification can be undercut, whether they support some more general principle, and whether that principle is independently defensible. (This is one of the tasks taken up in this chapter.)

A second way to defend a necessitarian requirement is to accommodate it within a principled theory of welfare goods. In particular, it fits well with *experientialist* theories, which hold that all welfare goods are experiential in nature (like felt pleasantness), such that an individual’s level of welfare is fully determined by their subjective experience.¹⁵ If all welfare goods are fundamentally experiential, then only conscious beings can have them. But experientialist theories are notoriously controversial. According to *extra-*

¹¹ I suspect that something seeming to be the case has a distinctive phenomenology—it *feels right*, it *feels true*. This is acknowledged in some accounts of the nature of intuitions (Huemer 2001; Chudnoff 2013), though my argument will not depend on it.

¹² Siewert (1998), Kriegel (2019, forthcoming), Chalmers (2022).

¹³ Kagan (2019, 28).

¹⁴ Harman (2003, 184), Kagan (2019, 31).

¹⁵ Advocates of experientialism include Kraut (2018), van der Deijl (2019, 2021), and most hedonists, though some hedonists understand pleasure as a pro-attitude, leaving open whether pleasure so understood requires phenomenology.

experientialist theories of welfare, an individual's level of welfare can vary without any difference in their subjective experience because there are welfare goods that are not experiential in nature.¹⁶ To many, for example, it seems better for them to know the real world than to be deceived by a simulation; better to write the book they have always wanted to write than merely to feel as if they had written it. Perhaps all such goods presuppose consciousness, but further argument is needed to establish an entailment of this kind.¹⁷

Theories of mind offer a third strategy for defending Necessitarianism. For if any mental state requires consciousness, and any welfare subject requires a mind, then only conscious beings can be welfare subjects.¹⁸ The first premise, however, is highly contentious. Leading theories of mental states define beliefs and desires by their functional and dispositional roles with respect to sensory inputs, behavioural outputs, and other mental states.¹⁹ On these views, consciousness might not be essential to mentality, and, in principle, a non-conscious AI could be said to desire the achievement of its goals, or to know how to translate a language.

Because all three routes are controversial, the necessitarian requirement seems to lack solid support, whereas *Non-Necessitarianism*—the view that being a conscious entity is *not* necessary for being a welfare subject—is becoming increasingly popular.²⁰ This may come with revisionary implications:²¹

¹⁶ Prominent extra-experientialist theorists include Nozick (1974), Finnis (1979), Hurka (1993), Bradford (2015), and Lin (2021).

¹⁷ I examine this view in Section 6.2.

¹⁸ For the view that all mental states have a conscious component, see Strawson (1994) and Georgalis (2006).

¹⁹ Stalnaker (1984), Smith (1987), Marcus (1990), Dennett (1991), Fodor (1998).

²⁰ Carruthers (1999), Levy (2014, 2024), Kagan (2019), Bradford (2022), Shepherd (2023), Mogensen (2025a), Goldstein and Kirk-Giannini (2025).

²¹ Some forms of Non-Necessitarianism can avoid at least some of these implications. Levy (2014, 2024), for example, argues that access consciousness—the availability of information for reasoning, verbal report, and the control of action—can suffice for morally relevant interests even in the absence of phenomenal consciousness. In the real world, most cases of access consciousness presumably coincide with phenomenal consciousness, so an access-consciousness

Rejecting [Necessitarianism] may result in an expansion of the population of welfare subjects. It may include sophisticated robots or even Roombas, some plants, and possibly other surprising things... Assuming that welfare generates normative reasons, it turns out that we may have reason to, say, make the world a better place by giving robots more knowledge and satisfied desires, or letting the weeds in your yard flourish.

Bradford (2022, 913)

How should necessitarians defend their view? An incredulous stare doesn't accomplish much. It would be more satisfying—and indeed more persuasive—to develop a general necessitarian argument that doesn't hinge on any particularly controversial philosophical theory, or hinges on it as little as possible. You might think there is none. That's why some have taken the necessitarian requirement as a basic datum that cannot be justified on further grounds,²² while others have dismissed it as a groundless prejudice, on a par with the speciesism Peter Singer (1977) criticised.²³ We would simply have reached bedrock. But I don't think we have. Perhaps we may never be able to give anything like a conclusive proof of Necessitarianism, but we can give some reasons to believe it. That's what I'll do in this chapter.

I'll defend the necessitarian requirement with a new epistemic argument. My argument proceeds in two steps. First, I argue that the introspection of certain experiences seems to provide *prima facie*, defeasible justification for the belief that things can go well or badly for a subject at a given time only if it is conscious at that time. Necessitarianism follows from this. Second, I contend that there is no defeating evidence against this

criterion would still exclude plants and very simple robots. But it is worth noting that this form of Non-Necessitarianism would still have other revisionary implications for the treatment of more sophisticated AIs: those that lack evidence of phenomenal consciousness but show evidence of significant access consciousness.

²² van der Deijl (2021, 192–193).

²³ Bradford (2022, 914).

introspectively justified belief: neither from apparently opposing intuitions, nor from wellbeing theory, nor from philosophy of mind. Since in the absence of defeating evidence *prima facie* justification should be taken as all-things-considered justification, I conclude that we ought to be necessitarians.

2. Introspection and Justification

There is a way to justify the indispensability of consciousness to welfare subjectivity. We need only to appeal to the special epistemic access we have to our own ongoing felt mental life: to look inside ourselves and *introspect* certain kinds of experiences.

Our inward window already supports many other beliefs. Suppose, for example, that my eyes are closed and I feel an unbearable, intense ache. If I introspect this feeling, I seem to have strong justification to believe I'm in pain. Presumably, it is the subjective character of my introspected experience—the presentness of the felt pain to my mind—that provides the justification for this belief. Suppose further that I possess the concept of 'intrinsic badness-for.' Upon introspection, I find myself strongly inclined to believe that being in pain is intrinsically bad for me.²⁴ What supports this belief is an *introspection-sensitive intuition*: an intuition based on how it feels to have a particular experience to the subject who introspects it.²⁵

The justificatory force of introspection-sensitive intuitions seems legitimate, strong, and mundane. Few claims in moral philosophy are less controversial than that pain is to some degree intrinsically bad for us.²⁶ Think

²⁴ In cases of morphine analgesia, prefrontal lobotomy, and pain asymbolia (Grahek 2007), some pains are reported not to hurt or feel unpleasant. Even if it were accurate to describe such states as pains, and even if such pains were not intrinsically bad for us, it could still be true that at least those pains that do hurt or feel unpleasant are in fact intrinsically bad for us.

²⁵ On a more radical view, introspection makes us directly acquainted with the intrinsic prudential badness of pain because this evaluative property is *constituted* by the phenomenal qualities of pain (Mendola 1990; de Lazari-Radek and Singer 2014; Sinhababu 2024; Dimitrov 2024). While I have some sympathy for this view, I leave open whether any such constitutive relation underlies our introspection-sensitive intuitions.

²⁶ For some resistance, see Delon (2024).

of quantified propositions such as ‘all pains that feel bad like this are intrinsically bad for me,’ or modal propositions such as ‘any pain that feels bad like this is necessarily intrinsically bad for me.’ Arguably, such propositions seem true in virtue of what it is like for us to have a pain that feels bad ‘like this,’ as we can appreciate first-personally through introspection.²⁷ I’ll argue that the necessitarian requirement has the same solid foundation.

In making my argument, I’ll assume that introspection-sensitive intuitions provide *prima facie justification* for beliefs. A *justification* for a belief, here, is a justification for taking something to be true; and it is *prima facie* if it justifies the belief unless and until there are reasons for doubt (*defeaters*). Prima facie justification, that is, constitutes *all-things-considered justification* in the absence of defeating evidence. Consider again the familiar case of pain: introspection gives me strong reason to believe that pain is bad for me, and if there is no reason for doubt, it seems I should believe it overall. (Some understand all-things-considered justification as knowledge. But I’ll remain neutral on the relationship between justification and knowledge since it is controversial.)

It’s worth stressing that, in my view, introspection-sensitive intuitions provide *defeasible* justification. Perhaps the most attentive introspection of the clearest phenomenology is infallible, immune to even the subtlest deceptions of sceptical scenarios. Perhaps I cannot be mistaken in believing that I’m in pain if I clearly seem to be introspecting pain. But I could still be wrong about the way I understand and conceptualise my own felt experience. I could, for example, be wrong about its intrinsic prudential badness. Historically, many have professed to feel the will of God inside them. Let me at least allow the possibility of a conceptual mistake.

²⁷ Kahane (2010) argues that anyone who doesn’t know how pain feels isn’t in a position to know that pain is bad. The thesis here is weaker. I’m only claiming that our introspective phenomenology of pain can play a justifying role in our acquiring the belief that pain is bad, not that this is the only legitimate way to acquire that belief.

We can systematise the above premises within some form of *phenomenal conservatism*.²⁸ On this framework, the very experience of φ 's *seeming* to be the case provides *prima facie* justification for believing φ in virtue of what it feels like to have that experience. More generally, my argument will be compatible with any theory on which introspection-sensitive intuitions are sources of (at least) *prima facie* justification. *Reliabilism* could be such a theory. If the justification of a belief depends on the reliability of its belief-formation process, introspection-sensitive intuitions might have special justificatory force insofar as they tend to be especially reliable (I'll return to this later).²⁹ What I'll say will also be consistent with *metaethical experientialism*: the view that we can obtain ethical knowledge in virtue of how certain phenomenal states feel, such as knowledge of the badness of pain.³⁰ Indeed, were my argument concerned with knowledge, it would conclude that we can know that consciousness is necessary for welfare subjectivity in virtue of how it feels for us to have certain experiences.

3. The Argument from the Night (Pt. 1)

The following argument differs from earlier intuition-based arguments for Necessitarianism in two respects. First, it rests not on a single data point but on a pattern across many intuitions, and is therefore less vulnerable to the worry that it places too much weight on a single idiosyncratic case. Second, these intuitions are based on features of mundane conscious experiences with which each of us is directly acquainted, and so the argument can appeal to the epistemic authority of introspection. What experiences, then, can give us

²⁸ Huemer (2001), Pryor (2000).

²⁹ Sinhababu (2024) defends hedonism by appealing to the putative reliability of introspection. But it seems to me an open question whether introspection inclines us to value conscious experience solely in virtue of its pleasantness, or also in virtue of other aspects of its phenomenology, such as its novelty, emotional significance, or aesthetic value (van der Deijl 2019).

³⁰ Lee (forthcoming).

reason to believe that only conscious beings are welfare subjects? I propose two: the experience of *losing* consciousness and the experience of *regaining* it.

3.1 Losing Consciousness

Consciousness doesn't stretch unbroken from birth to death. We all know it can be lost after a blow to the head, general anaesthesia, blackout drinking, or simply falling into a dreamless sleep.³¹ How does that feel? We cannot experience the lack of consciousness, but we can experience its *fading*: those transitional states between conscious and non-conscious episodes. And that fading has distinctive phenomenal qualities. Colours, sounds, and feelings grow murky and muffled. Our subjective point of view is reduced to a vague awareness in the dark, its borders getting fuzzier and fuzzier, only to disappear entirely. We can speak of 'blacking out' as a metaphor for this experience, since the light seems to turn off before our eyes. But we cannot literally see darkness in a dreamless sleep. Even the darkness shrinks inwards until any experience is gone.

Now, the fading of consciousness into dreamless sleep isn't perfectly gradual. It often begins with a flurry of hypnagogic imagery: flashes most often visual, but sometimes auditory, tactile, or kinaesthetic. And then, suddenly, everything blacks out. So we may never have in the course of our lives a single continuous experience through which we can introspect the gradual fading of conscious phenomenology from its paradigmatic richness down to its complete loss. What we can do is mentally reconstruct such an experience. Consider any sufficiently comprehensive set of experiences from our own lives in which the gradable features of consciousness—such as richness, intensity, or colourfulness—vary from very high to very low degrees. We can then arrange them in a sequence, where each point corresponds to an experience we have

³¹ This is the default view in the science of consciousness in sleep. Although some states currently classified as dreamless sleep may be minimally conscious, most researchers agree that at least a subset involve 'periods in which consciousness is lost' (Windt 2020, 12).

directly introspected, and each experience scores lower with respect to the relevant features than the last, so that, over the sequence, those features score lower and lower until they vanish entirely.³² Then, if some bare phenomenal property should remain once every gradable feature has reached its lowest score, we can extend the reconstruction one step further, to the point at which even that property has faded and nothing conscious is left.

As we lose consciousness over this reconstructed sequence, how well off we can be at a given moment—our *momentary* wellbeing—seems to vary considerably. Consider the moment consciousness is maximally diminished, when we are teetering on the brink of a lightless abyss. A barely conscious life doesn't seem worth much. More precisely, it is widely intuitive that if one is *minimally conscious*—in the sense that the gradable features of one's conscious experience, such as richness and intensity, are reduced to a bare minimum—then one's life at that time can be, at best, *barely* worth living, regardless of what else it contains. Think how little money, if any, you'd spend to add a few extra days of minimally conscious existence to your life. On those days your loved ones might visit, you might retain your knowledge, perhaps even achieve a few things. But from your own point of view, all you get are distant echoes, barely audible from the depths of the darkness you're in. This doesn't look like a great opportunity for wellbeing. And the intuition doesn't stop there. When we reach the bottom of this sequence and our subjective perspective vanishes entirely into the dark, it also seems that the resulting period of life cannot be worth living at all. From our own point of view, we don't even seem to be there. So how could things be going well or badly for us, at that time?

The intuition can also be reformulated in terms of *lifetime* wellbeing, and some may find this version especially persuasive. Instead of reconstructing

³² The reconstruction need not be a strict ordering in which each experience scores lower than its predecessor on all the relevant features at once. Perhaps the actual experiences of a typical human life cannot be arranged into such a complete sequence. Even so, I suspect we can mentally reconstruct a rough trajectory in which the relevant phenomenal features fade away as an overall trend, using introspected experiences as waypoints.

a sequence of experiences in which the features of consciousness are gradually lost within a single life, we can reconstruct a *series of entire lives*, each lived by a subject whose experiences score lower than those of the preceding subject with respect to the gradable features of consciousness. Over the series, these subjects' experiences shrink from life to life until consciousness itself is gone. It seems strongly intuitive that a life lived throughout at the most minimal state of consciousness is, as a whole, and regardless of what else it contains, at best barely worth living for the one who lives it. And at the bottom of the series, where even the last spark of experience has gone out and there is only perspectiveless darkness, no subject seems to be there at all—certainly not from that individual's own point of view. So how could such a life, taken as a whole, go well or badly for them?

Let me say more about the intuitions I've elicited. Together, their contents form the following intuitive evaluative judgment: the closer we come to losing consciousness, the closer our momentary or lifetime wellbeing level seems to come to zero, so that, were consciousness lost entirely, we could no longer have any positive or negative wellbeing at that time or over that life. This intuitive judgment is grounded in introspected experience, namely, in what it's like to approach the loss of consciousness itself. Indeed, although the sequence of experiences or lives is reconstructed in thought, every phenomenal feature used to construct it has been directly introspected and contributes to the judgment's intuitive force.

A few clarifications are in order. This is not an intuition about the instrumental value of losing consciousness. It's an intuition about the intrinsic wellbeing value that can be accrued when consciousness is nearly lost or entirely gone. Of course, as we fall asleep we may feel warm and comfortable, or cold and tired. But these are merely *contingent* properties of the phenomenology of losing consciousness, and our reconstructed sequence should abstract away from them, since they are in no way plausible indicators of the properties of non-conscious states themselves. Socrates appears to have

made the basic mistake of confusing these two sets of properties when he supposed that, if death amounts to an eternal dreamless sleep, having such a never-ending night of pleasantly restful sleep would be good for us.³³ Let's not make the same mistake.

Another potentially confounding element is that in the reconstructed sequence of experiences approaching the loss of consciousness we lose many other things besides phenomenology. We lose the ability to use informational content for reasoning, action, and speech—our *access* consciousness.³⁴ We also lose the ability to act as we choose, and the will to resist what is being done to us—our *agency*. But among all these losses, what grounds our introspection-sensitive intuition is, I think, our vanishing phenomenology. To see this, imagine losing phenomenal consciousness while preserving everything else as much as possible. To the extent that I continue to be an access-conscious agent, I continue to behave in various ways. But to the extent that I continue to lose my phenomenology, I continue to feel relevantly like any other time I'm losing consciousness. That is, my subjective perspective still seems to fade into some kind of darkness, and it still seems to me that a barely conscious life—a life devoid of rich and intense phenomenal qualities—would be, at best, barely worth living for me.

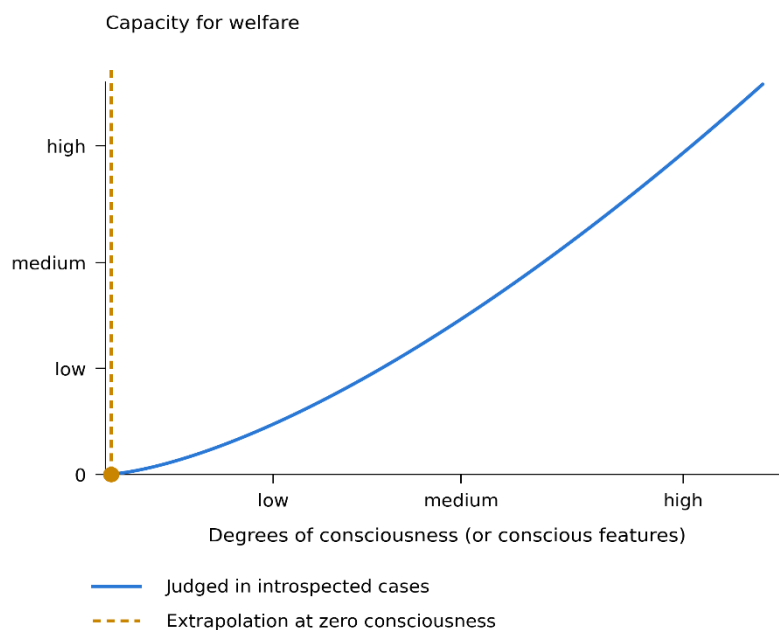
Some might report not sharing this intuition. I'm inclined to think that they may have been led to a particular interpretation of their experience by unjustified worries about its ethical implications. They might worry that the intuition implies that we lack moral status overnight, that nothing bad can happen to us during a non-conscious episode, or that experientialism about wellbeing is true. But, as we'll see along the way, the pre-theoretical intuition is free from these worries. And pre-theoretical intuitions of this sort should be our starting point; theories should, and will, come later. After all, theories of welfare are constructed out of intuitions, not the other way around.

³³ Plato's *Apology* (40d–e).

³⁴ Block (1995).

There is admittedly a jump from the intuitive judgment that a minimally conscious life has at best very little positive or negative welfare value to the intuitive judgment that a non-conscious life has none at all. It may be hard to see how the latter judgment could be supported by introspection, since we obviously cannot introspect non-conscious states.³⁵ In my view, the jump is justified by a *trend-based inductive inference*, that is, an inference from a pattern observed in experienced cases to a conclusion about unexperienced cases. The trend is this: *the more we lose consciousness, the less we seem capable of accruing welfare goods and bads at that time or over that life, regardless of what else that time or life contains*. When consciousness is minimal, the level of welfare available to us seems neutral or nearly so. We can extrapolate from this trend that without consciousness we would lack the capacity for any positive or negative welfare level. It is in this sense that our intuitive judgment about the non-conscious case can rightly be said to be grounded in the phenomenology of *approaching* the loss of consciousness.

The trend is represented schematically below (the curve's particular shape is illustrative only).



³⁵ See Carruthers (1999, 478) and Shepherd (2023) for versions of this objection.

Trend-based induction is a familiar and respectable form of inference, widely relied on in science and medicine, and presumably preserves the *prima facie* justification provided by introspection-sensitive intuitions.³⁶ Of course, the extrapolation that *every* bit of welfare is lost at zero consciousness is, like any induction, open to counterevidence. One might argue, for instance, that in the trend the capacity for welfare only comes close to zero without exactly reaching it, so that very small amounts of welfare, positive or negative, remain possible even once consciousness is lost.³⁷ But two things can be said in defence of the zero-point extrapolation.

First, it seems the most natural extrapolation from the trend. As consciousness fades, we can intuit a progressive decline in the capacity for welfare at every point in the sequence, and at the last such point that capacity *already* seems to be at the zero point, or extremely close to it. The most straightforward continuation of this trend is that the capacity for welfare reaches zero when consciousness is lost. To insist instead that some residual capacity for welfare remains precisely where the introspected trend runs out is to posit a more specific and less natural curve. The zero-point extrapolation, then, has considerations of theoretical simplicity on its side.

Of course, the correct curve may well be a more specific and less natural one. But, and this is my second point, any specific curve-shape hypothesis requires independent motivation. In science, for example, we posit an *asymptotic* curve when we have reason to think that the trend has a limiting structure: some mechanism or constraint that prevents the value from ever reaching zero and allows it only to come indefinitely close. In the case at issue,

³⁶ Inferences of this kind transmit justification: if one is *prima facie* justified in believing the premises of a good inductive argument, and competently infers its conclusion, one is *prima facie* justified in believing that conclusion. Of course, the extrapolated conclusion may be less securely justified than its premises. But my claim is only that the belief retains *prima facie* justification, and whether it has that status is not itself a matter of degree. The question, then, is whether the belief should stand in the absence of defeaters, and I see no reason to think that a trend-based induction would undermine that status.

³⁷ Such a curve would be compatible with Bradford's (2022) claim that non-conscious individuals can accrue only very small amounts of wellbeing.

we would be entitled to posit a curve that doesn't reach zero only if we had reason to think that something could ground welfare even in the absence of consciousness, that is, if we had reason to postulate the existence of welfare goods with no conscious component. But, as I'll later argue, it seems that we have no such reason (Section 6.2). Until such a reason is given, 'it might not reach exactly zero' is a mere formal possibility. And mere possibility doesn't defeat *prima facie* justification.³⁸

In my graph, I represent consciousness as a matter of degree.³⁹ Don't read too much into that—everyone agrees that either consciousness itself comes in degrees, or at least many of its features do. If it's the gradual loss of consciousness itself that seems to correlate necessarily with the gradual loss of our capacity for welfare, we can extrapolate from this trend that consciousness itself is necessary to be able to accrue welfare, and thus to be a welfare subject. If, instead, consciousness is an all-or-nothing affair, and it's only the richness, intensity, and other features of consciousness that gradually degrade to zero, we can nonetheless extrapolate a revised version of the necessitarian requirement:

Feature Necessitarianism. Having certain conscious features is necessary for being a welfare subject.⁴⁰

Perhaps not all gradable features of consciousness are relevant to welfare. The saturation of an experience, for instance, may be irrelevant. What matters here is that reductions in at least *some* conscious features, such as richness and intensity, seem to correlate necessarily with our capacity for welfare approaching the zero point. We can extrapolate from this trend that having at

³⁸ The bare possibility that the curve might take some other form is not defeating evidence. It is the observation that induction is not deduction, which is true of every inductive conclusion. If mere possibility defeated *prima facie* justification, perhaps nothing would ever be *prima facie* justified.

³⁹ On degrees of consciousness, see Lee (2023a).

⁴⁰ Feature Necessitarianism can be expressed in both the state and capacity versions of Necessitarianism distinguished in Section 1 and discussed in Section 5.

least *some* conscious features is necessary to be able to accrue welfare goods and bads, and thus to be a welfare subject.

3.2 Regaining Consciousness

There is another way to appreciate the prudential significance of consciousness through introspection. Consider the experience of suddenly regaining consciousness after a short afternoon nap, jolted awake by a loud alarm. We are all familiar with the impression of nothingness that precedes our abrupt awakening. Where was I? *When I have a dreamless sleep it feels like I time-travel.* I disappear, and then I'm back. The same happens under general anaesthesia: no time seems to have passed between the moment I was counting down and the moment I wake up. Later I may realise that I haven't actually time-travelled; the clock kept ticking without gaps. But the felt duration that this interval had for me—my *subjective time*—was compressed to zero.

The experience of waking can then give us a first-hand grasp of the notion of subjective time. For present purposes, I'll understand subjective time as the *felt* duration that a period has for a given subject, the kind of duration that most obviously seems to go missing when I black out in dreamless sleep and to return when I wake up. I can then wonder: how well could my life go for me without it?

Admittedly, I cannot directly introspect a life without subjective time and form an intuitive judgment based on how it feels. What I can do, again, is extrapolate such a judgment from a broader trend in our introspection-sensitive intuitions. This time, instead of gradually reducing richness, intensity, or the overall degree of consciousness, we should imagine gradually reducing to zero only *the rate of subjective time*. By this rate I mean the felt duration generated per unit of objective time: how long a fixed objective interval feels to a given subject. As this rate approaches zero, we approach a life without subjective time, a life for which any objective interval has no felt duration at all. Of course, we may never directly experience such a continuous

decline in the rate of subjective time. But over the course of our lives we have many experiences of differing subjective durations, and we can arrange them in a mentally reconstructed sequence in which each point corresponds to a rate of subjective time we have directly introspected.

Let me consider, then, a life with good things spread evenly throughout its objective extension in time, but which feels shorter and shorter to me as I live it. An interval of the same objective length first feels like a decade, then a year, then a month, a day, an hour, a second, and so on. There are still friends, achievements, love, and joy, but the subjective time to appreciate any of it is compressed over and over again, until it's compressed to zero. How well could things go for me over each period in the sequence?

Intuitively, there is a downward trend, all else being equal. The briefer a period feels to me, the less wellbeing I seem able to accrue over it. When my rate of subjective time is minimal, and every event feels as if it lasts only the most fleeting instant, life seems, at best, barely worth living. There may be friends, achievements, love, and joy, but from my own subjective perspective it all happens so quickly that it seems as if it (almost) never did. All these potentially good things slip through my fingers and leave me very, very little. We can extrapolate from this trend that without a subjectively felt duration of time things could not go well for me at all. I am therefore *prima facie* justified in believing that being a welfare subject requires having a subjective time, and thus being conscious.⁴¹

Note again that the sequence can also be reconstructed as a succession of entire lives with differing rates of subjective time, each feeling shorter to the one living it than the last. All else being equal, the lifetime wellbeing of each

⁴¹ Following Boethius, one might think that God exists outside of time, and so perhaps lacks subjective time. Would this preclude God from being a welfare subject? Traditionally, God is not understood as the sort of being that can be made better or worse off; God is not vulnerable or imperfect in that way. Perhaps God could be *infinitely* well off. But it seems hard to get an intuitive grasp on the idea of a God who lacks subjective time and yet is infinitely well off, if this is intelligible at all. Therefore the case of God should at best be taken as an unreliable counterexample.

subject seems to shrink over the sequence. A life that as a whole feels no longer than the most fleeting instant seems, at best, barely worth living for the one who lives it, regardless of what it contains. We can extrapolate that a life without any felt duration cannot be good or bad for one. Having such a life is prudentially no different from never having been born, since, from the subject's own point of view, it feels as if it never happened at all.⁴²

I realise that there may be other ways of understanding subjective time that do not require conscious phenomenology, and perhaps the lives we are imagining could have a subjective time in some non-phenomenal sense. But insofar as these forms of subjective time make no difference to how long a period *feels* to the subject, I doubt they will seem intuitively relevant to welfare. Perhaps, for example, a non-conscious AI could be said to have its own subjective time in the sense that each interval has a *judged duration* for it—a length of time it reports or believes has passed. But mere judgments about the duration of a period are not directly relevant to welfare. Someone who lived five happy days might mistakenly believe she had lived five centuries of happiness; yet if the felt duration was in fact only five days, this mistaken belief doesn't seem to make her better off.

4. The Argument from the Night (Pt. 2)

Let us take stock. We have a range of experiences that *prima facie* justify belief in Necessitarianism. One is the experience of losing consciousness itself. Using gradations of conscious features we have directly introspected, we can mentally reconstruct a sequence of experiences approaching a minimal state of consciousness: a state in which conscious features such as richness, intensity, or subjective time are reduced to a bare minimum. The more we lose consciousness in this way, the less welfare we seem capable of accruing at the

⁴² One might think that a being without subjective time could live a life that feels as if it never ends; a single moment of subjectivity that never moves forward for it and is endlessly appreciated in experience. But it seems to me that such an experience would still have a subjective duration, simply an infinitely extended one.

time. At the minimally conscious point in the sequence, any life which felt like this seems, at the time, at best barely worth living for the one living it. It is then from this trend in our intuitive judgments that we can justifiably extrapolate the following belief:

Night. For any individual S and time T, if S has no conscious experience at T, then S cannot have a positive or negative momentary wellbeing level at T.

Night establishes a necessary condition for all welfare states: *no consciousness, no welfare*. This is distinct from experientialism about welfare. Once I'm conscious, good- or bad-making features that don't affect my felt experience could affect how well my life goes. Experience, then, may not be all that matters to wellbeing. But without it, nothing does. At any time between the moment at which consciousness completely fades away and the moment at which experience breaks through the darkness, nothing could make my life go well or badly.

Why should an individual be unable to have a positive or negative momentary wellbeing level at a time at which it has no experience? The best explanation is that it cannot accrue welfare goods or bads at that time, that is, the things that can make a life go well or badly at that time. The alternative—that goods and bads accrue but necessarily balance out—attributes to the individual a preordained harmony that would remain entirely unexplained. I'll therefore assume that *Night* licenses the following formulation: an individual can accrue welfare goods or bads at T only if it has conscious experience at T.

In some ways, *Night* is more fundamental than Necessitarianism. It fits the introspective data more closely, has wider implications, and, I suspect, is what makes Necessitarianism true in the first place: if nothing can be intrinsically good or bad for a being at a time when it lacks phenomenal experience, then nothing can be intrinsically good or bad for a being that lacks even the capacity for phenomenal experience, and thus never has it.

But *Night* is also more radical than Necessitarianism. The necessitarian requirement leaves open whether one can be well off at a time at which one is unconscious. *Night* rules this out. Yet intuitively there are ways in which we can be harmed or benefited while unconscious—say, if our partner cheats on us, frustrating our desire for a faithful relationship. Whether these intuitions should count as defeaters for *Night* is an important question, which I address later in Section 6.2. But two things are worth noting now.

First, what approaching the loss of consciousness feels like from the inside puts pressure on the idea that we can accrue welfare goods and bads in its absence. We should be open to the possibility that the cheating itself is not what's bad for us, that what's bad, if anything, is something else. Perhaps what's bad is committing ourselves to the wrong kind of relationship *when we are conscious*, either because a relationship prone to infidelity violates a requirement on genuine love, or because it simply isn't the kind of relationship we want.

This leads to a second point. Even if extra-experiential welfare goods and bads can be triggered when we are not conscious, it remains an open question whether they can be accrued at that time. In the cheating case, it's not clear when our wellbeing level is affected and the harm is accrued—it may not be at the time that the cheating occurs. If it turns out that the harm is accrued when we are conscious, this would be consistent with *Night*. (Indeed, I'll later argue that extra-experientialist welfare theories can be understood in this way).

I expect that some theorists will remain unconvinced, especially those with the strong intuition that, at least in some cases, extra-experiential goods or bads are accrued *precisely* when they seem to occur, whether or not we are conscious at that time. This would give them reason to resist *Night*. It is worth noting, then, that the patterns in our introspection-sensitive intuitions can support the necessitarian requirement through a second route.

Recall that the losing-consciousness and losing-subjective-time sequences can each be reconstructed in two ways: as a sequence of experiences

within a single life, or as a sequence of whole lives. *Night* was extrapolated from the first sequence: from a trend in the *momentary* wellbeing levels that can be accrued by the same subject as they gradually move from full to minimal consciousness. But a parallel claim can be extrapolated from the second: from a trend in the *lifetime* wellbeing levels that can be accrued over whole lives, as each subject in the sequence experiences, at every point in its life, a lower degree of consciousness than the subject before it, until we reach a life lived at minimal consciousness throughout. Given that pattern, we can extrapolate:

Lifetime Night. For any individual S, if S has no conscious experience at any point in its lifetime, then S cannot have a positive or negative lifetime wellbeing level.

On *Lifetime Night*, a being unconscious throughout its life cannot have a life that as a whole goes well or badly for it. *Lifetime Night* is, in effect, *Night* applied at every moment of a life: if a being has no experience at any moment, from birth to death, then, by *Night*, it cannot have a positive or negative momentary wellbeing level at any point in that life. And so, on the assumption that lifetime wellbeing supervenes on momentary wellbeing, it cannot have a positive or negative lifetime wellbeing level either. But we need not accept *Night* in order to accept *Lifetime Night*. Even someone who blocks the extrapolation to *Night* still faces the extrapolation to *Lifetime Night*, since the latter rests on its own trend.

I take it that *Lifetime Night* can also be expressed as the thesis that any being lacking conscious experience throughout its lifetime cannot accrue welfare goods or bads. This best explains why such a being cannot have a positive or negative lifetime wellbeing level. We can thus derive a version of Necessitarianism:

Lifetime Necessitarianism. Being in a phenomenally conscious state at some point in one's life is necessary for being a welfare subject.

This captures the core motivation for Necessitarianism, since the requirement is primarily meant to exclude entities that are not conscious at any point in their existence from being welfare subjects, such as rocks, non-conscious robots, and the like. Lifetime Necessitarianism might diverge from Capacity Necessitarianism in a restricted range of cases, but the gap may be empty or negligible. Can a being have the capacity for consciousness and yet never exercise it across a whole life? Can a being that has been or will be conscious be said to lack that capacity at another time? If not, the two versions coincide extensionally.

Now that *Lifetime Night* is on the table, we might wonder whether there is any reason to keep the more radical *Night*. I think there is. Only from *Night* can we derive Capacity Necessitarianism, the most popular version of the necessitarian requirement. Moreover, without *Night* we leave Necessitarianism open to a troubling objection. If there could be beings who are not phenomenally conscious at a given time and yet are living lives that are going well for them at that time—beings who are having a good time in existence even though there is nothing it is like to be them—why, then, couldn't an entirely non-conscious being have that too? Perhaps only beings capable of consciousness can have welfare goods or bads, because all such goods have a conscious component. But if we can have these goods when we are not conscious, the connection between the conscious component and the good becomes less direct, needs further argument, and licenses some bizarre assertions (as I argue in Section 6.2).

From now on, I'll focus on *Night*. This is partly because it is an intuition I myself am inclined to share, and partly because if *Night* is true then *Lifetime Night* follows, but not vice versa. Moreover, if there are reasons to reject *Night*, they might not tell against the weaker thesis, *Lifetime Night*, which I can keep as a fallback position. And it's a *solid* fallback position. The intuition that a life minimally conscious at all times is, at best, barely worth living for the one who lives it seems as solid as an intuition can be in ethics—as solid as the intuition

that pain is bad, and at least as solid as the intuition that much is missing from a life plugged into a machine of merely simulated experiences. With that in hand, we can afford to pursue the bolder claim, *Night*.

5. The Argument from the Night (Pt. 3)

Let me now spell out the route from *Night* to Necessitarianism. Recall that Necessitarianism comes in different forms. On the state version, a being can be a welfare subject at T only if it is in a conscious state at T; on the capacity version, only if it has the capacity for consciousness at T. In principle, *Night* can support either. This is because the standard definition of a welfare subject—‘a being whose life *can* go well or badly for it’—is ambiguous. The modality of ‘can’ may be relative to the properties a being could have given the sort of being it is, or to the properties it could have under more specific conditions, such as at a particular time or in a particular situation. Consider pain. In general, James is the kind of being that can feel pain—he is capable of pain. Yet it may also be true that James cannot feel pain while in a dreamless sleep. He lacks the capacity for pain at that particular time. The same ambiguity extends to welfare subjectivity. In general, James is the sort of being that is capable of accruing welfare goods or bads—he is a welfare subject in a *kind-relative* sense.⁴³ Yet he cannot do so while in a dreamless sleep; he lacks the capacity to accrue welfare goods or bads at that particular time, and so is not a welfare subject in a *time-relative* sense.

We can now make sense of the distinction between State Necessitarianism and Capacity Necessitarianism. Given *Night*, we should be state necessitarians in the time-relative sense and capacity necessitarians in the kind-relative sense. *Night* implies not only that (i) only a being in a conscious

⁴³ ‘Kind’ here should not be understood as a biological kind. An anencephalic human infant might not be the kind of being that can accrue welfare goods, even though humans generally are. Rather, an individual is the ‘kind’ of being that is generally capable of accruing welfare goods if it itself possesses the kind of properties that make it capable of accruing such goods, such as, perhaps, being capable of enjoyment or desire.

state can accrue welfare goods or bads at a given time, but also that (ii) only a being with the capacity for consciousness can be the sort of being that is generally capable of accruing welfare goods or bads. For even in the deepest slumber, a being with the capacity for consciousness may remain *at that time* the kind of being that would be capable of accruing welfare goods and bads *at other times*, whereas a being lacking the capacity for consciousness is the kind of being that could not accrue welfare goods or bads *at any time*. In a slogan, ‘*no consciousness, no welfare*’ here essentially means ‘*no conscious states, no welfare goods or bads*.’ This in turn reads either as ‘*no conscious states, no welfare subjectivity*’ or as ‘*no capacity for conscious states, no welfare subjectivity*,’ depending on whether welfare subjectivity is understood in its time-relative or kind-relative sense.

I suspect both notions are useful for different purposes. When it comes to attributing moral status to an individual, we are concerned with its kind-relative welfare subjectivity (even during a dreamless sleep, we remain the sort of being that is generally capable of having welfare goods and bads, and so we have interests that matter morally, at least as dispositions). By contrast, when it comes to determining whether an individual can be harmed or benefited at a particular time by a particular event, we are concerned with its time-relative welfare subjectivity.

These two notions of welfare subjectivity seem morally significant in a variety of contexts. But there is another understanding of welfare subjectivity that I wish to resist, precisely because it seems to lack such significance. On this view, any being with a welfare level counts as a welfare subject, even if that level is *neutral*.⁴⁴ This view threatens to block the move from *Night* to Necessitarianism, since *Night* allows non-conscious beings to have a neutral level of welfare. Indeed, as we approach a minimal state of consciousness, we seem to approach a necessarily neutral level of welfare, and there is no clear reason to extrapolate from this trend that the neutral level itself is lost once

⁴⁴ van der Deijl (2021, 2025), Lee (2025), and Kriegel (forthcoming).

consciousness is lost. So, absent further argument, combining this understanding of welfare subjectivity with *Night* doesn't yield the conclusion that only conscious beings are welfare subjects.⁴⁵

The problem with a neutral-level-inclusive definition of welfare subjectivity is that it undermines the role this concept is meant to play in ethical theory. For when it comes to ethical questions, what really matters is whether a being has welfare goods or bads, or is at least capable of having them. Consider a being that necessarily cannot accrue welfare goods or bads but nonetheless has a neutral level of wellbeing—perhaps a being that can only see the faintest brightness without feeling any pleasure or pain or forming any beliefs or desires. Call this being *Zero*.⁴⁶ What is the point of regarding *Zero* as a welfare subject? It's hard to see how *Zero* could have moral status. Nor is there any need to include *Zero* in our population axiology, since either it makes no difference to the aggregate value of a population, or, if it does, that seems like an implausible feature of the theory.⁴⁷ There is nothing at stake in the case of *Zero*, morally speaking. Because I care about having a robust link from welfare subjectivity to normative ethics, I restrict the status of welfare subject to those

⁴⁵ Advocates of this definition of welfare subjectivity typically argue that a being lacking even the capacity for consciousness doesn't seem to be a subject for whom things can go well or badly, or even *neutrally*. By contrast, Lin (manuscript) argues that such entities can have a necessarily neutral welfare level, given that it seems legitimate to make welfare comparisons between us and them. For what it is worth, I'm inclined to agree. A happy human being seems better off than a table. It is not indeterminate, unanswerable, or puzzling whether I am better off than a mere thing—I just am. I have a good life; it does not. So the mere thing must have a determinate welfare level; otherwise I could not compare our welfare levels and conclude that I am better off. The way to exclude such entities from being welfare subjects is to define a welfare subject as an entity capable of accruing welfare goods or bads, which mere things are not.

⁴⁶ *Zero* is introduced by Lee (2025).

⁴⁷ When we calculate the average welfare in a world, one might think that we must take into account beings like *Zero* (Lee 2025). This would have implications for non-separable population axiologies such as averaging—but odd ones. For example, a world populated by terribly unhappy beings could be made significantly better simply by adding sufficiently many *Zeros*. But how could hell be improved by the addition of beings that lack even the capacity for a life worth living? I suspect that these population axiologies should be rejected or revised so that they factor in only the welfare of welfare subjects.

beings that can have welfare goods or bads, excluding *Zero* and the like.⁴⁸ (For an extended version of this argument, see Chapter 2, Section 5.5).

Now that we have provided some reason to believe that only conscious beings can accrue welfare goods or bads, and that only beings capable of doing so are welfare subjects, we have all the materials for an epistemic argument for Necessitarianism. Call this the *Argument from the Night*:

(P1) We have a range of introspection-sensitive intuitions that provide us with prima facie justification for believing *Night*: for any individual S and time T, if S has no conscious experience at T, then S cannot accrue welfare goods or bads at T.

(P2) In the absence of undefeated defeaters, prima facie justification constitutes all-things-considered justification.

(P3) There are no undefeated defeaters for *Night*.

(C1) Therefore, we are all-things-considered justified in believing *Night*.
(P1–P3)

(P4) If *Night* is true, then S is, at T, the sort of individual that can accrue welfare goods or bads only if S is, at T, the sort of individual that can have conscious experience—that is, only if S has, at T, the capacity for conscious experience.

(P5) S is a welfare subject at T just in case S is, at T, the sort of individual that can accrue welfare goods or bads. (*kind-relative welfare subjectivity*)

⁴⁸ There is then a discontinuity in the kinds of entities that are welfare subjects: all beings that are capable only of good or bad lives are welfare subjects, yet all beings that are capable only of neutral lives are not. But this discontinuity is not arbitrary as Lee (2025) suggests. In fact, it reflects a genuine discontinuity in moral significance.

(C2) Therefore, we are all-things-considered justified in believing that S is a welfare subject at T only if S has, at T, the capacity for conscious experience. (*Capacity Necessitarianism*) (C1–P5; justification transmits across entailment)

Since my argument relies on defeasible justification, accepting it doesn't guarantee that *Night* and Necessitarianism are true. We could be wrong. Perhaps the internal workings of a non-conscious mind bear welfare value, but we lack the cognitive capacities to grasp even the possibility of such value. Who knows? It's time to consider candidate defeaters.

6. Defending Night

So far, I've tried to justify Necessitarianism by appeal to evidence that should, absent defeaters, persuade any rational person: we can appreciate that consciousness is necessary for welfare subjectivity by observing its fading, or that certain conscious features are necessary by observing their fading. In principle, however, one may lack this evidence. There could be beings who never dreamlessly sleep, who never lose consciousness; beings incapable of even *imagining* the fading of consciousness, just as a sighted person who had never closed her eyes might be unable to imagine what it is like to be blind. I can't argue with them, just as I can't argue with someone sceptical of the intrinsic badness of pain. The best I can do is wait for them to experience pain themselves. If they cannot, I will have a hard time. But I'm fine with that. There are countless necessary truths that we can learn only thanks to contingent facts about our cognition.

The intuitions that support *Night* are relevantly like the intuition that pain is intrinsically bad. Both are based on how certain experiences feel upon introspection. Since beliefs formed on this basis require no further justification beyond experience itself, the advocate of *Night* mainly bears the burden of showing that the arguments against it are weak. To this end, I'll examine two

kinds of defeaters: (i) evidence of malfunction in the introspective mechanism responsible for its seeming to me that *Night* is the case (*undercutting defeaters*), and (ii) evidence in favour of not-*Night* (*rebutting defeaters*).

6.1 Undercutting Defeaters

Let us consider five candidate defeaters that, if successful, would undercut our reasons for believing *Night*.

I

I've been assuming that our stream of consciousness can be broken. But what if it merely fails to be stored in memory and reported in behaviour? What if, even in the deepest slumber, we remain conscious in some basic, rudimentary sense?⁴⁹

The worry finds its most radical expression in *panpsychism*. On this view, the experience of approaching what we would naturally describe as the *loss* of consciousness would be nothing more than the experience of approaching a *more basic* form of consciousness. For on the panpsychist view, *everything* that exists at a time is conscious at that time, so, if we continue to exist, we must remain conscious in some basic sense. It may then become unclear how such an experience could support the belief that only conscious entities are welfare subjects (except in a trivial sense, since all entities would be conscious). But it can. As long as the panpsychist agrees that, when we approach the more rudimentary form of consciousness, we can have only a neutral or near-neutral momentary wellbeing level, we can still run a trend-based induction. Indeed, as certain gradable conscious features (intensity, richness, and the like) fade away, we observe a downward trend in how well our

⁴⁹ Some theorists take this idea very seriously. Thompson (2014, Ch. 8) draws on the Indian and Tibetan philosophical traditions to argue that subtle forms of consciousness may persist in dreamless sleep.

life can go at that time. From this, we can extrapolate that having these features is necessary for being a welfare subject (Feature Necessitarianism).

One such feature of experience may be its *valence*: what makes it feel good or bad. I see no good reason to believe that all things—chairs, stars, comatose beings—have experiences that feel pleasant or unpleasant. And I'm open to the possibility that consciousness may be necessary for welfare subjectivity because some particular conscious feature such as valence is necessary.⁵⁰ However, I will not argue for this here, as I'm also attracted to the necessitarian requirement in its most general form, and restricting it to valenced conscious features would commit me to a needlessly strong thesis for my purposes.

II

I've argued that by introspecting the experience of losing consciousness we can intuitively appreciate that we need to be conscious to be welfare subjects. But it may seem unclear how we can justifiably extrapolate a universal, necessary condition from our own particular case. After all, AIs and robots could have non-conscious minds vastly different from our own. Why should our intuitive judgment extend to them?

In my view, the scope of the intuition *prima facie* includes every possible being by virtue of its character. It is just like the intuition that *any* pain that feels bad like *this* is necessarily intrinsically bad for its subject. The intuition was never about my own life in particular. Nor was it ever about a particular kind of human-like non-conscious mentality. After all we might have all kinds of non-conscious mental states when we are minimally conscious, and we don't know which ones we have, if any. So the intuition is not that *this* or *that* non-conscious mental state doesn't matter. It is that *whatever* mental states we might have during a minimally conscious life, they

⁵⁰ In fact, this is what I'll argue later in Chapter 2.

could not make that life better than barely worth living, for it seems that *any* life which felt like this could be, at best, barely worth living. That's the kind of introspection-sensitive intuition at issue: one that purports to apply universally and necessarily to all possible beings. And the mere fact that an intuition with necessary and universal content is generated by a particular mind like our own is not itself a reason to reject it. If that were enough to undercut justification, mathematical intuitions would be in trouble too.

III

One might worry that introspection-sensitive intuitions are generally unreliable belief-forming processes. Schwitzgebel (2008) is surely right that it is hard to form fine-grained introspective judgments about the phenomenology of emotions or mental processes. Introspection might be especially unreliable when experiences are so rich and complex that we can mischaracterise their details. But even the most sceptical philosophers should admit that we can reliably introspect strong, vivid pain. More generally, I'm not aware of any convincing evidence that introspection-sensitive intuitions are unreliable when they register *high-stakes variations in welfare levels* sensitive to variations in phenomenology, as in cases of severe pain (towards the negative), intense pleasure (towards the positive), or severe anhedonia (towards the neutral).⁵¹ Schwitzgebel may be right that we are often uncertain if we ask ourselves, 'Do you really feel bad while doing the dishes? Are you happier going to a wedding or to a restaurant with your family?' (2008, 250). But these are difficult cases, as they concern small variations in wellbeing levels. Let's instead ask ourselves which is better: a nice restaurant or torture, a joyful life or a life of anhedonia. We have no doubt. I restrict my argument to the apparent reliability of those introspection-sensitive intuitions that register high-stakes variations in welfare levels—cases such as severe pain, intense pleasure, severe

⁵¹ Anhedonia is a condition characterised by a severely diminished ability to feel pleasure or enjoyment.

anhedonia, or *the loss of consciousness itself*, which intuitively looks like a total loss of welfare in virtue of how it feels.

IV

One might think that the very existence of extra-experiential goods undermines the authority of introspection-sensitive intuitions about welfare. After all, insofar as introspection acquaints us with goods of any sort, they must be goods that belong to conscious experience; insofar as there are goods that do not supervene on conscious experience, introspection will be blind to them. So don't we have reason to doubt that introspection-sensitive intuitions can give us knowledge of the full range of goods that are available to us in a minimally conscious state?

I recognise that intuitions grounded in introspected experience may not give us exhaustive knowledge of how well or badly our life goes. All I claim is that they give us *prima facie*, *defeasible* justification for belief. If from the inside my life seems to be going well, I should take my experience to give me *prima facie* justification to believe that my life is indeed going well, even though I know that in principle I could be wrong, since I cannot rule out the possibility that I am hallucinating in an experience machine. I shouldn't give up my belief just because I cannot exclude such possibilities. I should give it up only if I have positive evidence that some goods or bads insensitive to my experience are in fact obtaining.

So the question is whether there is any positive evidence in tension with *Night*. To settle it, we should use *every tool* we have for investigating wellbeing, not just introspection. But when we do, it turns out that we need not postulate that a life without consciousness could go well or badly in order to explain the relevant evidence. We can accommodate all the evidence for *Night* and for extra-experientialism within a single theoretical package, at no extra cost (Section 6.2). None of this establishes beyond doubt that a subject cannot accrue goods or bads at times at which it is unconscious. But we should not

posit more welfare goods or bads than we have reason to. Parsimony, then, favours *Night*.

V

Perhaps it's unsurprising that a *conscious* being would form the belief that only *conscious* beings are welfare subjects through a *conscious* mechanism like introspection. Is my argument biased? I don't think it is. Feeling exhausted and thinking 'I'm exhausted' is not a bias. Feeling happy and thinking 'I'm happy' is not a bias. Rather, introspection-sensitive intuitions provide non-arbitrary reasons for belief.⁵² These beliefs are not unexamined prejudices either: I'll now examine their justification against opposing evidence, and explain why such evidence fails to rebut *Night*.

6.2 Rebutting Defeaters

Without evidence undercutting *Night*, non-necessitarians need a *rebutting* defeater: either (i) undefeated intuitions that independently justify not-*Night*, or (ii) a powerful theory that rules out *Night*, such as a theory of welfare. The problem, as I see it, is that they have neither. I'll argue that non-necessitarian intuitions can be easily defeated, and that any theory of welfare can be made compatible with *Night* without significant extra costs. I'll therefore conclude that we are all-things-considered justified in believing *Night*, as it is better justified than its denial. For imagine that I am in pain and this seems bad for me, but a theory T says otherwise. Suppose that (i) I have no undefeated intuitive justification for believing T, and (ii) I can explain the original motivations for T in other ways—some of them compatible with my being in pain and with this pain being bad for me—without incurring significant extra costs. Clearly, I should still believe that I am in pain and that this is bad for me;

⁵² See Lee (2022) for an extended defence of the view that restricting welfare interests to conscious beings is disanalogous to discrimination based on species membership, race, or sex.

it would be irrational to give up my belief in this epistemic situation. The same will be true of *Night*.

Of course, making certain theories compatible with *Night* may involve *some* costs. But, as I'm going to show, it is doubtful whether these costs are unacceptably high, or even costs at all. And when it comes to theory choice, I take the desideratum of accommodating introspectively justified beliefs to override small differences in theoretical costs between competing theories.⁵³

With this in mind, let us consider four rebutting defeaters.

I

Can we directly intuit that some non-conscious entities are welfare subjects? Kagan (2019, 28) proposes a thought experiment to show we can. Imagine meeting a group of sophisticated alien robots. You somehow learn that they lack consciousness, so you capture one to cut it open and inspect its inner mechanisms. However—and this is the crucial point—the other robots beg you not to kill their companion. They insist that they are not mere things and claim to have strong desires, hopes, and ambitions for their own lives, just as a person would. What should you do?

Kagan has 'no doubt whatsoever' that it would be wrong to kill one of these robots. He is inclined to attribute moral status to them because the sophistication of their behaviour is sufficient proof that they really do have desires, hopes, and ambitions. And since it would be odd for a being to have these attributes without being a subject of harms and benefits, I take Kagan to have a non-necessitarian intuition about welfare subjectivity as well.⁵⁴

⁵³ It remains an open question whether this desideratum can override *very significant* differences in theoretical costs. Perhaps I should favour a very complicated theory that accommodates the seeming that 'I am in pain and this is bad for me' over a much simpler theory that doesn't. But nothing in my argument requires this stronger claim.

⁵⁴ Elsewhere in the book, Kagan seems to accept that non-conscious beings may have welfare (2019, 33–35).

Speaking personally, I don't share this intuition. The lives of these creatures are relevantly like a permanent dreamless sleep. And intuitively, no being in such a state seems capable of the kind of desires, hopes, or ambitions whose fulfilment could make a life go well. Moreover, non-necessitarian intuitions seem easily undercut: these robots look as if they are conscious, so we can expect our intuitions about them to be unreliable. We cannot simply flip a mental switch and suppress our normal intuitive responses when faced with something that looks like a conscious being in extreme distress. That said, I can still see why it could be wrong to destroy these robots for no good reason. But this has nothing to do with their welfare, and everything to do with the cultivation of a morally problematic character or attitude in the person who destroys them. After all, it also seems intuitively wrong to fantasise all day about killing a purely fictional character and cutting them open to examine their insides, yet few would say that a fictional being can suffer genuine harm.⁵⁵

II

It has been claimed that plants have wellbeing regardless of whether they are conscious because it is intuitive to say that things can be 'good for' or 'bad for' them. However, the claim that 'good for' in this context means 'contributing to welfare' needs to be argued for. Only philosophers use 'good for' strictly in this way. In ordinary language 'good for' usually means 'conducive to health,' which is why I misuse the language if I say that eating sweets is good for me. Although eating sweets might raise my momentary wellbeing level, the assertion sounds false because eating sweets isn't healthy. In the plant case there is no reason to understand 'good for' in the welfare sense rather than the health sense, since the intuitive appeal of saying that something is 'good for' a fern is completely accounted for by what makes ferns healthy. Without further

⁵⁵ See my *Double-Standard Moralism* (2023) for an argument according to which fictive imaginings can be morally criticised.

argument, then, appeals to talk of what is ‘good for’ plants establish only that plants can be healthy or diseased, which is uncontroversial.

III

The simplest combination of extra-experientialist theories of welfare and non-phenomenal theories of mental states seems to conflict with *Night*. Call this the *Combination Argument*. It runs as follows: (i) there are extra-experiential welfare goods, such as knowledge or desire-fulfilment; (ii) non-conscious entities can have mental states, such as beliefs or desires, that typically generate such goods; and (iii) such mental states can generate welfare goods whether or not they are conscious. It follows that non-conscious entities can have welfare goods.⁵⁶

These theories can be construed to block this argument without significant extra costs. The simple solution is to accept that every extra-experiential welfare good has multiple components, including at least an inward mental state and an outward state of the world, and that one of these components is always a state of consciousness. This explains why only conscious beings can have such goods.⁵⁷ On such a *conscious-component theory of welfare*, what is intrinsically good for me is having *conscious knowledge*, that is, knowledge I have while I’m phenomenally conscious; fulfilling *conscious desires*, desires I’ve had while I’ve been phenomenally conscious; completing *conscious projects*, projects I’ve pursued while I’ve been phenomenally conscious; and so on for any other candidate welfare good, such as *conscious achievements*, *conscious friendship*, or *conscious love*.

⁵⁶ For versions of this argument, see Levy (2014), Bradford (2022), and Shepherd (2023). A recent version is developed by Goldstein and Kirk-Giannini (2025), who argue that some existing artificial language agents could be welfare subjects.

⁵⁷ Something like this is proposed by Lin (2021) and Kriegel (forthcoming). Lin argues that the truth of Necessitarianism can be explained by the fact that all welfare goods require consciousness in virtue of their natures. Kriegel considers two theories of wellbeing on which the only goods are experiences and *experiential organic unities*, namely the coming together of certain worldly components and at least one conscious component.

There is nothing particularly revisionary about this view. We are not modifying our theory of mental states; we are only clarifying what our theory of welfare goods may have held *all along*. The requirement that an intrinsically valuable state have a conscious component might simply have been too obvious to be worth saying. After all, our ordinary conception of desires, knowledge, projects, achievements, friendship, and love is intimately intertwined with phenomenal experience in many ways.

Some may insist that the conscious-component theory of welfare is ad hoc.⁵⁸ But it is not. Extra-experientialism has been primarily motivated by intuitions about cases such as the *experience machine*.⁵⁹ Imagine being plugged into a machine that provides you with wonderful conscious simulations of a life that never exists outside your mind. Comparing this simulated life to the one you have, you might conclude there is more to wellbeing than experience. But if you do, presumably you have formed that intuition by contemplating how much better it is to have *conscious* knowledge, *conscious* friendships, and *conscious* achievements than to be conscious but lack those things. For you were asked to compare your ordinary life with one plugged into the machine, and you are conscious in both cases. Consider the analogue without any conscious component: a machine that vastly increases one's *non-conscious* knowledge, *non-conscious* achievements, or the fulfilment of one's *non-conscious* desires. People wouldn't find such a machine particularly appealing, certainly not appealing enough to explain the rise of a view as popular as extra-experientialism about wellbeing. If so, the conscious-component view fits the intuitive data that motivated extra-experientialism better than its no-

⁵⁸ van der Deijl (2021) argues that extra-experientalist theories of welfare can be made compatible with Necessitarianism only by adding the ad hoc constraint that only conscious beings can have extra-experiential goods. He therefore concludes that we should accept experientialism about welfare, which offers a simple, non-ad hoc explanation of why only conscious beings are welfare subjects. I agree that experientialism is simpler, and that simplicity counts for something. But I deny that every necessitarian version of extra-experientialism is unacceptably ad hoc.

⁵⁹ Nozick (1974).

conscious-component rival does. Rather than being ad hoc, then, it is independently motivated by the very intuitive data that made extra-experientialism attractive in the first place.

Let me address two further worries. First, one might demand a principled explanation of why consciousness is a necessary component of welfare goods. Yet extra-experientialists don't typically provide a principled explanation of why welfare goods have a worldly component. The intuitive appeal of the idea is enough. Why should we treat the conscious component differently? Second, one might object that conscious-component extra-experientialism is less simple than its no-conscious-component rivals, since it adds a component. But the gain in faithfulness to the original intuitive data arguably compensates for this loss in simplicity. All things considered, then, the conscious-component view doesn't seem significantly more costly.

Since extra-experientialist theories of welfare can be revised to block the Combination Argument without significant extra costs, this argument doesn't rationally require us to give up *Night*.

IV

I have argued that extra-experientialism can be construed so that every welfare good has a conscious component, and that this construal is faithful to its original motivation. But it seems coherent to think both (i) that having knowledge, friendships, fulfilled desires, and the like requires being conscious at *some* time, and (ii) that these goods can benefit their subjects at times at which those subjects are not conscious. Were this the best way to understand the conscious-component view, *Night* would face a defeater.

Note that this line of objection already grants that every welfare good has a conscious component. *Lifetime Night* is therefore immune to it, since on the conscious-component view an entity that is never conscious could not have welfare goods. So even if the defeater succeeded, there would still be another route to Necessitarianism. But I think *Night* can be defended as well.

The natural way to make the conscious-component view compatible with *Night* is to restrict the momentary accrual of extra-experiential goods to times at which the subject is conscious. Welfare goods may be *triggered* when I am unconscious, in the sense that many of the worldly conditions necessary for the obtaining of the good may be met at those times (for example, a paper I wrote might be accepted while I am under general anaesthesia). But their goodness can be *accrued* at a time only if I am conscious at that time, since being conscious at a time is a further subject-relative condition that must be met for the good to accrue at that time. This condition is *Night*: for any individual S and time T, S can accrue welfare goods or bads at T only if S has conscious experience at T.

Again, I think we can reconstruct our theories of welfare to accommodate this condition while respecting their original motivation. If every welfare good has a conscious component, but something can benefit me when I lack consciousness, the connection between the conscious component and the welfare good becomes less direct. It becomes less clear in what sense consciousness is part of what makes the good good for me *then*. By contrast, we can naturally spell out the idea that every welfare good has a conscious component if we accept the requirement that a subject can have such goods at a given time only if the subject is conscious at that time. This is a plausible specification of the conscious-component view, and the one that keeps consciousness most tightly bound to the good of which it is supposed to be a component.

Admittedly, this is not the only possible construal of the view. One might argue that the conscious component belongs to the good's *history*. Consciousness may be needed to form a desire or belief, or develop a friendship, even if the resulting goods can benefit their subject at a later time, when the subject is no longer conscious. However, a diachronic interpretation of the conscious-component view would require further argument: it would need to specify, exactly, what the conscious component of each good is, if not

the subject's being conscious when the good is accrued. And even if such a story could be told, there are two more reasons why the synchronic interpretation of the conscious component seems to me especially promising.

The first is that it avoids licensing assertions that sound bizarre. Many good things can happen while I am unconscious. But it would be strange to say that these things are intrinsically good for me *at that time*. If they were, it would be appropriate to say things like:

1. 'The best hour of my week was the hour I spent under anaesthesia, because my book received a glowing review then.'
2. 'My friends often came to visit while I was in a coma, entirely unconscious. That period of my life wasn't so bad after all.'
3. 'The best part of the conference, for me, was the afternoon I missed because I had fainted, since that was when people praised my work.'

These statements sound odd. To be clear, it does seem good that my book received an appreciative review, or that I have caring friends. But intuitively, these things do not make my life go well *at the very time at which I am unconscious*. We should take the oddness of these statements as evidence that the goodness of such things is accrued at some other time, if at any particular time at all.

One might object that it sounds equally odd to say that things which happened when I was unconscious can benefit me at other times, when I am conscious. The second reason to accept the synchronic interpretation of the conscious-component view is that, on reflection, this isn't so strange after all. In fact, the very nature of extra-experiential welfare goods can be reconstructed to accommodate this synchronic condition at no extra cost. We need not understand such goods as free-floating worldly facts: the achievement itself, the desire-fulfilment itself, the relationship itself, and so forth. We can instead

understand them as subject-relative properties: *as ways in which my life as a conscious subject can go*. What is good for me, on this view, is consciously living as someone who is objectively successful, loved, knowledgeable, fulfilled, or connected to others. These properties come in degrees. At a time at which I am conscious, I can be *more* or *less* successful depending on worldly events—whether, for instance, I was awarded a prize overnight at a committee meeting on the other side of the world—and so my life can go more or less well for me at that time.

Redescribing our theories of welfare goods in this way respects the original motivation for extra-experientialism while explaining why I can accrue such goods at a time only if I am conscious at that time, and it seems to do so without significant extra cost. Still, one possible cost is worth considering. It may seem plausible that if I now desire my daughter to have a good life, and she has such a life only after I die, this is still good for me. Or that it is good for me if, after my death, my life's projects are finally successful. *Night* might seem to rule out these putative *posthumous benefits*, at least if the following independently plausible principle is true:

Completion. S can receive a benefit at T only if, at T, all the conditions necessary for S to receive that benefit have been met.⁶⁰

If being conscious is necessary for accruing any good at a given time, and if I cannot be benefited at a time before all the necessary conditions on my receiving a benefit are met, then I cannot be benefited at a time by events that occur after my death. At any time before my death, the relevant worldly conditions—the obtaining of what I desired, the success of my projects—have not yet been met, whereas at any time after my death, I am not conscious. Thus there is no time at which all the necessary conditions on my receiving the benefit are satisfied.

⁶⁰ Introduced and motivated by Lin (2017).

But *Night* can resist this challenge too. On some views of posthumous benefit, events after my death do not benefit me at any particular time.⁶¹ They simply make it true, *atemporally*, that my life as a whole has been more fulfilled or successful than it would otherwise have been. This is neither in tension with *Night*, since I am not benefited at any particular time at which I am unconscious, nor with *Completion*, since *Completion* constrains only benefits received at a time. This does raise a question: what is the conscious component of an atemporal benefit? When a benefit is accrued at a time, the conscious component is my being conscious at the time I accrue the good. When a benefit is accrued atemporally, to a whole life, I suggest that the conscious component is that life's having been the life of a conscious subject (that is, one conscious at some point in that life).

Of course, one might argue that in cases of posthumous benefits I am benefited at a particular time. If one rejects *Completion* and takes that time to fall within my lifetime, then that view can be made compatible with *Night* without extra costs, since the benefit can be accrued at a time at which I was conscious (such as when I had the conscious desire, or when I was consciously working on my projects).⁶² If instead I am benefited at some post-mortem time, the view faces a well-known difficulty: there seems to be no suitable bearer of the benefit, since the formerly living subject no longer exists at that time, if not as a corpse.⁶³ So either the views that accept posthumous benefits are already consistent with *Night*, or they come with serious costs of their own.

The general lesson is that extra-experientialist theories of welfare can be made compatible with *Night* on the assumption that every welfare good can be accrued at a time only if the subject is conscious at that time. These theories

⁶¹ Boonin (2019, 153) grants that posthumous harms and benefits might make a subject's life as a whole go better or worse atemporally, without being accrued at any particular time. But he primarily defends the view that such harms and benefits are accrued at times at which the person was still alive. For an explicitly atemporal account, see Johansson (2012, 476).

⁶² For the view that desire-fulfilment benefits one at the time of the desire, even when what is desired obtains only at a later time, see Dorsey (2013).

⁶³ Feinberg (1984, 80), Boonin (2019, 116–121). See also Chapter 3 (Section 1).

are not merely logically possible but also plausible, since they neither betray the original motivation for extra-experientialism nor incur significant extra costs.

7. Conclusion

On what grounds do we have reason to believe that only conscious beings are welfare subjects? I have argued that this belief is introspectively justified. When we look inside ourselves and experience the fading, the loss, and the regaining of consciousness, it seems as if we are also experiencing the fading, the loss, and the regaining of what makes life worth living. We sink into a dark, perspectiveless life that lacks any subjective duration for the one who lives it; a life that, intuitively, cannot go well or badly. Perhaps we are victims of some error, but the route by which we have formed this belief seems so far clear, smooth, and free of defeaters. We have no choice but to take it.

Chapter 2:

The Mind that Matters

What kind of mind is necessary and sufficient for moral status—for having interests that matter morally in and of themselves? This chapter defends Affective Sentientism, the view that moral status requires the capacity for affective experiences such as pleasure, pain, and emotion. I argue that affective consciousness is what makes an entity a welfare subject, and that all and only welfare subjects have moral status. The connection between affect and welfare is systematic: I show that every major candidate welfare good has an affective component, often because such goods require the capacity to care. I also argue that even if the properties of being phenomenally conscious, being autonomous, or having a neutral level of welfare could come apart from affective consciousness, none of them could ground moral status in its absence. I conclude by exploring the implications for the moral status of animals and artificial intelligence systems.

1. Introduction

Imagine yourself in a dreamless sleep. You don't think, see, or feel anything. There is only an impenetrable darkness, if even that. The prospect of this state, however, doesn't worry you: you know that you will later regain *phenomenal consciousness*—that colourful stream of experience that feels like something from the inside. But suppose you lacked the capacity for consciousness, had never had it, and never would. Then, one might think, your existence would not be importantly different from that of a rock or a fungus. It would seem *worthless*—not in the sense of being useless, but in the sense that nothing could be intrinsically good or bad for you, and others would see no reason to care about you for your own sake.⁶⁴ Indeed, for the one who lives it, a life immersed in that impenetrable darkness might seem just as good as not existing at all.

The underlying intuition is that the capacity for some form of consciousness seems *necessary* for having interests that matter morally in and of themselves—that is, for being an entity with *moral status*. Some form of consciousness also seems *sufficient* for such status, since at least all beings capable of experiencing suffering plausibly deserve moral consideration. *Sentientism* is the view that captures these two intuitions: the capacity for consciousness, in some form, is both necessary and sufficient for moral status.

Let me take Sentientism for granted. Though contested, it is arguably the dominant view about the grounds of moral status in contemporary ethics.⁶⁵ In this chapter, I focus on its precise form. As defined above, Sentientism takes no stance on which kind of conscious experience is required for moral status. But under the umbrella of consciousness there are many different experiential properties: from the euphoric character of a hearty laugh

⁶⁴ I assume that fungi lack intrinsic moral significance insofar as they are not conscious. Proponents of biocentrism will disagree; for a critique of biocentrism, see Basl (2019).

⁶⁵ Defenders of Sentientism include Bentham (1907), Singer (1977, 1993), DeGrazia (1996), Siewert (1998), McMahan (2002), Korsgaard (2018), Shepherd (2018), Powell and Mikhalevich (2020), van der Deijl (2021, 2025), Lin (2021), Lee (2022, 2025), Chalmers (2022), Roelofs (2023), Perry (2024), Smithies (2026), and Kriegel (forthcoming).

to the felt emptiness of loneliness, and from the striving character of experiences of effort to the felt spark of understanding something new. If these features of consciousness can come apart, so that some remain while others are permanently lost, then what would have to be missing for a life to lack moral status? Would it have to be consciousness altogether, or would it be enough to lose a particular kind of consciousness? Which is the mind that matters?

The traditional answer is *affective consciousness*: conscious experience that feels good or bad. As I use the term, affective experience encompasses states with a positive or negative hedonic valence, such as pleasure and suffering, as well as emotions and desires that involve such hedonic states. This subset of phenomenal consciousness is sometimes called *sentience*.⁶⁶ But since there are other ways of spelling out Sentientism, I will use the term *Affective Sentientism* for the view that all and only beings capable of affective consciousness have moral status.⁶⁷

Let us remain neutral on how the notion of capacity is to be understood. It is clear why the capacity for affect suffices for moral status: if there are moral reasons to care about others at all, it is hard to deny that the possibility of causing them to feel joy or suffering warrants at least *some* moral consideration. What is less clear is why affect is necessary—why we should restrict our circle of moral concern to *affective beings*, that is, beings capable of affective consciousness.

Indeed, a recent trend has been a move towards more inclusive forms of Sentientism.⁶⁸ Some have proposed that all phenomenally conscious beings have moral status, and that even a simple capacity to experience light would suffice for it. I call this view *Phenomenal Sentientism*.⁶⁹ Others maintain that

⁶⁶ Singer (1977, 8–9) and Birch et al. (2021, 7). Others, such as Kagan (2019, 13), use ‘sentience’ to mean phenomenal consciousness simpliciter.

⁶⁷ Bentham (1907), Singer (1977), DeGrazia (1996), Shepherd (2018), Powell and Mikhalevich (2020), and Smithies (2026).

⁶⁸ Chalmers (2022, forthcoming), Birch (2022), Roelofs (2023), Shepherd (2024), van der Deijl (2025), and Kriegel (forthcoming).

⁶⁹ Chalmers (forthcoming) is inclined to accept this view.

moral status requires any kind of consciousness that can provide a being with a reason for action—a view I call *Motivational Sentientism*.⁷⁰

The general reason to reject an affective restriction on moral status is that it has consequences many find unsettling. Consider a *Philosophical Vulcan*—a highly sophisticated, phenomenally conscious creature without the capacity for affective feelings.⁷¹ The idea is inspired by the fictional species in *Star Trek*. But while *Star Trek*'s Vulcans occasionally feel emotions, Philosophical Vulcans are by stipulation completely incapable of positive or negative affect (hereafter, this is what I mean by 'Vulcans'). Even so, they make plans, form intentions, and act autonomously, guided by their own rational self-government. They also get all that non-affective consciousness has to offer: experiences of colours, freedom, intentionality, understanding, and perhaps even puzzlement, selflessness, and agency. They possess all the potentially morally relevant capacities an entity can have without affect. It is therefore natural to ask: Wouldn't they have their own interests and ambitions? Wouldn't it seem wrong to kill one for no good reason, even if many others could take its place? Many philosophers have the strong intuition that Vulcans would deserve moral consideration more or less as we do.⁷² This is because, despite their lack of affect,

[Vulcans] might still have a rich conscious life, with multimodal sensory experiences and a stream of conscious thought about all sorts of complex issues ... They may want to advance science ... and to help those around them ... They experience no pleasure when anticipating or achieving these goals, but they value and pursue the goals all the same.

Chalmers (2022, 336)

⁷⁰ Roelofs (2023).

⁷¹ The Vulcan case is introduced by Chalmers (2022, 335–7). For earlier examples, see Carruthers's *Phenumb* (1999, 478–80) and Crisp's anhedonic being (2006, 122).

⁷² Chalmers (2022, forthcoming), Birch (2022), Roelofs (2023), Mogensen (2025b), van der Deijl (2025).

It may turn out that such creatures are not merely conceivable but actually possible. In existing humans and animals, phenomenal and affective consciousness tend to go together. In the future, however, what is known about conscious experience could be operationalised in a functional description of conscious processing that abstracts from hedonic experience, and could then be implemented in artificial systems. These would be *Artificial Vulcans*: sophisticated silicon-based creatures who could reflect on the intricate movements of the wind, sense an approaching storm through their sensors, and form the intention to seek shelter—all without ever feeling the unpleasantness of a shiver. Should we include such beings in our circle of moral concern when the time comes?

In this chapter, I examine what it would be like to live a sophisticated conscious life entirely devoid of affective experience. After sustained argument and reflection, I conclude that Affective Sentientism has been right all along. It is the Vulcan who matters less than we might have thought.

My defence of Affective Sentientism rests on some more fundamental claims about the nature of welfare. I show that every major candidate welfare good and bad is best understood as having an affective component. In particular, many such goods and bads require the capacity to care, which essentially involves affective experience. Because all and only beings capable of welfare goods and bads are welfare subjects, I argue that affective consciousness is what makes an entity a welfare subject. And because all and only welfare subjects have moral status, I argue that affective consciousness is what makes an entity a bearer of moral status. Along the way, I explain why the properties of being phenomenally conscious, being autonomous, and having a neutral level of welfare could not ground moral status in the absence of affective consciousness. I close by exploring the implications for the moral status of animals and artificial intelligence systems.

2. The Argument for Affective Sentientism

Why think that affective consciousness grounds moral status? The argument I defend is this:

(P1) All and only beings whose interests matter morally in and of themselves have moral status.

(P2) All and only welfare subjects have interests that matter morally in this way.

(C1) Therefore, all and only welfare subjects have moral status.

(P3) All and only affective beings are welfare subjects.

(C2) Therefore, all and only affective beings have moral status.

Let me consider each premise in turn. The interest-based definition of moral status (P1) is the one sentientists typically endorse, and I'll take it for granted. On this view, moral status is distinct from intrinsic moral value.⁷³ Cultural symbols, languages, and artworks might be intrinsically morally valuable—there may be moral reasons not to damage or destroy them, independently of any effects on anything else—but they do not have moral status, because they do not have interests. An entity has interests just in case there are ways in which things can go better or worse for the entity itself, either relative to what makes its life go well or relative to what matters to it. Cultural symbols, languages, and artworks have neither lives that can go well for them nor things that matter to them, and so no moral reasons concerning them can be grounded in their interests.

Now, which entities have interests that matter morally in and of themselves, and thus moral status? A good answer is given by (P2): all and only entities with *welfare interests*, that is, interests in receiving benefits and

⁷³ Kamm (2007, 229), Chalmers (2022, 335–9).

avoiding harms. Such entities are *welfare subjects*: entities capable of having positive or negative components of *welfare*, or *wellbeing*, or whatever we call the property that makes a life good or bad *for* the subject who lives it. As I'll argue later, welfare subjects are entities capable of more than a merely neutral level of welfare. They can be bearers of *welfare goods and bads*: the states or conditions that are intrinsically good or bad *for* a subject and that constitute the positive and negative components of welfare (standard examples of which are pleasure and suffering, the fulfilment and frustration of desires, knowledge and ignorance, and virtue and vice).

It is nearly universally accepted that if you have welfare interests, you have interests that matter morally in and of themselves, and thus that *all* welfare subjects have moral status.⁷⁴ For the moment, I also take for granted that *only* welfare subjects do. There may be morally significant interests that are not grounded in welfare. Later I'll argue, however, that only welfare subjects can have them, because the same properties required for welfare interests are also required for non-welfare interests to matter morally. Hence, welfare subjectivity is both necessary and sufficient for moral status.

Given these modest assumptions (P1 and P2), the question of which entities have moral status reduces to the question of which entities are welfare subjects. The central thesis of this chapter is that *all and only affective beings are welfare subjects* (P3). Given the definition of welfare subjects introduced earlier, this is equivalent to saying that all and only affective beings can be bearers of welfare goods or bads.

Since (P3) is the most contentious premise in my argument for Affective Sentientism, I'll devote most of what follows to defending it. I'll argue that every plausible welfare good or bad requires the capacity for

⁷⁴ For discussion, see Lee (2022). Some theorists, such as Basl (2019, 161–82) and Kagan (2019, 31), deny that all welfare subjects have moral status, on the grounds that plants and mindless artefacts can be benefited or harmed in ways that do not matter morally. But I take it that when a flower is cut or a car is damaged, we speak of 'harm' only in a loose or figurative sense, not in the technical sense relevant to welfare. See also Chapter 1 (Section 6.2).

affective experience, and thus that only affective beings are welfare subjects.⁷⁵ Moreover, because the capacity for affective experience entails the capacity to have at least one welfare good or bad, every affective being is a welfare subject. It seems, indeed, that at least one of the following is true: that pleasure and suffering themselves are welfare goods and bads, or that any being capable of experiencing them can desire to have or avoid them in such a way that satisfying or frustrating those desires constitutes a welfare good or bad. All and only affective beings, then, are welfare subjects.

I call this thesis

The Affect-Welfare Link. The capacity for affective consciousness is necessary and sufficient for being able to have welfare goods or bads.

I call the deeper thesis that best explains this link

The Affective Thesis. The capacity for affective consciousness is what makes an entity a welfare subject.

In making my argument, I'll follow a methodology that, I think, should guide the search for any welfare good. Call it the *Welfare Good Extraction Method*. We begin by identifying something—say, hanging out with friends—that seems intuitively good for the person whose life includes it. We then try to isolate what matters from other factors and, from patterns in these intuitive judgments, we generalise that a certain thing—say, friendship—is itself a welfare good. But what happens if we strip away some property typically bound up with that candidate welfare good—say, the joy of seeing our friends? We cannot assume that what's left is still good. To do so would be to commit what I call the *Extraction Fallacy*: treating a modified version of ϕ as if it automatically inherits the welfare value of the original ϕ . There may be forms

⁷⁵ See Mogensen (2025b, 346–52) and Siewert (manuscript) for similar lines of argument, though the details differ in important ways.

of joyless friendship, but for all we know some shared joy may have been essential to its goodness. We therefore need to retest the intuitive plausibility of joyless friendship as a welfare good. Should it lose its appeal, we would have reason to think that the joy we removed from friendship was essential to its being a welfare good. That intuitive appeal was, after all, the reason we considered friendship a welfare good in the first place.⁷⁶

We could avoid the Extraction Fallacy if, whenever we formed the intuition that ϕ is good, we did so by grasping its *essence*: what is necessarily present in every instance of ϕ . But this is not how theorists typically form intuitions about welfare goods—partly because the essence of ϕ is often unclear (what is the essence of friendship?), and partly because our intuitions about the goodness of ϕ can be shaped by our ordinary conception of ϕ , which may include properties that are not strictly necessary for ϕ to be instantiated, yet are necessary for ϕ to count as a welfare good. That is, it is entirely possible that whenever we speak of ϕ as a welfare good, we really mean ϕ *plus* a set of properties that ordinarily come with ϕ . Our ordinary conception of friendship, for example, might include some form of shared joy, and it may be friendship so understood that we have in mind when we judge friendship to be a welfare good, regardless of whether friendship as such necessarily requires shared joy.

The Welfare Good Extraction Method then works as follows:

- I. Start from intuitive judgments about cases.
- II. From stable patterns in these intuitions, isolate what matters and extract candidate welfare goods.

⁷⁶ Versions of the Extraction Fallacy can be found in Levy (2014), Kagan (2019, 33), Bradford (2022), Shepherd (2023), and Goldstein and Kirk-Giannini (2025). Though these authors focus on phenomenal consciousness rather than affect specifically, the argumentative pattern is the same. They argue that welfare goods can be instantiated without some property they are typically bound up with, but either do not address whether what remains is still intuitively a welfare good, or their attempts to do so are, in my view, unsuccessful.

III. When a property typically accompanying a candidate welfare good is removed, test whether the original intuitions hold up.

IV. If they don't, the candidate welfare good should be understood as requiring that property.

Following this method, I'll argue that every welfare good has an affective component, because whenever we remove it from a candidate welfare good, that good loses its intuitive appeal.

Consider first pleasure and pain themselves. I remain neutral on whether these states are essentially attitudes, feelings, or both. In my view, however, for them to count as welfare goods or bads, they must involve pleasant or unpleasant conscious experience.⁷⁷ There may be a coherent notion of pleasure and pain without consciousness, perhaps understood as internal states that causally explain behaviours such as approach and avoidance in response to bodily benefit or damage. But without further argument, the existence of non-conscious hedonic states wouldn't entail an expansion of our welfare goods and bads. To assume otherwise would be to commit the Extraction Fallacy, since our concepts of pleasure and pain *as welfare goods and bads* may always involve conscious experience. Indeed, these are the pleasures and pains hedonists typically have in mind: conscious states that feel good or bad.⁷⁸ What's more, if we imagine having non-conscious hedonic states with *no effect at all* on our conscious experience, it doesn't seem that we would thereby be made intrinsically better or worse off. The lack of a clear intuition here should caution us against valuing pleasures and pains that do not involve consciousness, if any exist.

There are also candidate welfare goods whose goodness does not depend solely on their affective character, such as desire-fulfilment,

⁷⁷ Some pains, such as those in cases of pain asymbolia, are reported not to hurt (Grahek 2007). If these are pains at all, they may lack an affective component and so may not be the kind of pain that on my view is a welfare bad—namely, pain that feels unpleasant.

⁷⁸ Crisp (2006), Bradley (2009), Bramble (2016).

knowledge, virtue, love, friendship, and achievement. Although I'll discuss these goods extensively, it's helpful to outline my argument early on. I'll argue that even when these goods may not seem to rest on pleasure or suffering, an affective component is in fact essential to all of them in every instance. They require, at the very least, the subtler affective feelings involved in *caring about* someone or something.

The basic idea is that caring, in the relevant sense of the term, requires complex emotional states and dispositions partly constituted by episodes of affect. A central component of caring is a disposition to feel joy or satisfaction when the object of one's care is doing well, sadness or frustration when it is doing badly, fear when it is in danger, and relief when it turns out to be fine. Even when nothing actually happens to the object of one's care, one may be disposed to feel positive or negative affect *at the mere thought* of these possibilities. Some of these affective states may be absent without precluding care. But if all of them are missing, then, I'll argue, we are no longer talking about caring in the sense relevant to welfare.

I'm not sure whether caring is necessary for all welfare goods and bads. Perhaps it is possible to imagine a subject enjoying a stream of pleasant experiences despite lacking the psychological capacities needed to care about them in the relevant sense. A newborn infant may be an example: it may feel pure joy without being able to care about that joy. Yet such joy would still seem good for it. My claim, then, is that all welfare goods and bads require some affective component—either the complex affective states involved in caring or simpler hedonic states. And since at least one welfare good might not require caring, and the capacity for welfare goods suffices for morally significant interests, I'm inclined to believe that it is the capacity for affect, rather than caring, that grounds moral status.⁷⁹

⁷⁹ While Jaworska (2007) argues that caring grounds full moral status, or personhood, I remain neutral on that question. My concern here is only with which kind of mind is necessary and sufficient for moral status simpliciter.

The role I give to affect respects the idea that every basic component of welfare is appealing, engaging, and shouldn't leave one cold.⁸⁰ However, I do not rely on the intuitive appeal of such a *resonance constraint* on wellbeing as a ground for giving affect this role. Other theorists appeal to the requirement that what is good for someone shouldn't leave them cold, and argue that such engagement requires the heat of affective feelings.⁸¹ My argument is different, partly because it's not clear to me whether the resonance constraint *per se* is strongly intuitive as a basic principle about welfare, especially given the long-standing debate between subjectivist and objectivist theories of welfare. What strikes me as strongly intuitive is that affect is required for each particular candidate welfare good. It is from this intuitive data about particular goods that one should motivate a general resonance constraint, not the other way around.

It might be thought that we need a more unified explanation for the systematic correlation between affect and welfare goods. But for any given theory of welfare, the explanation is as unified as the theory allows it to be. If hedonism is true, and only one thing is a welfare good, then all welfare goods turn out to require affect for only one reason. If an objective-list theory is true, and more than one thing is a welfare good, then all welfare goods turn out to require affect for more than one reason. The right explanation need not be unified insofar as the basic nature of welfare is not unified. In fact, in that case the lack of unity would not be a flaw in my explanation, but a feature of the phenomenon I am explaining.

I begin by showing why desire-fulfilment requires affect, and then turn to other goods.

⁸⁰ Railton (1986, 9).

⁸¹ Fanciullo (2025a), Smithies (2026).

3. Vulcans, Desires, and Caring

Many think it is intrinsically good for us if our desires are *fulfilled*—if the world matches how we want it to be—and intrinsically bad if they are *frustrated*. To avoid counterexamples, however, wellbeing theorists tend to qualify which kinds of desires matter for welfare. I'll argue that the right kind of desire necessarily involves affect.

An initial question is whether desires are possible at all without affective consciousness. For us, desiring something always comes with a disposition to feel excitement or satisfaction if we have what we desire, or imagine having it, and a disposition to feel disappointment or even suffering if we do not have it, or imagine not having it. There might indeed be a necessary connection between having a desire and having such affective dispositions.⁸² But for present purposes, I'll grant that this connection need not hold, at least in alien minds. After all, some evidence suggests that the neural circuits of 'liking' (that is, enjoyment) can dissociate from those of 'wanting' in pathological cases.⁸³

What would an affectless desire look like? Vulcans might intentionally set goals for themselves by representing certain ends as positive ('to-be-pursued') or negative ('to-be-avoided') without any feeling of pleasure or displeasure. Vulcans, indeed, would seem capable of value judgments. They could judge a painting to be good by common aesthetic standards, even though the experience of it does not feel good to them. They might have the capacity for *cognitive valence* (thinking 'thumbs-up') without having that for *affective valence* (feeling 'thumbs-up'). And if they can judge a goal to be valuable, they may be motivated to pursue it for that very reason.⁸⁴ This would be an instance of *affectless desire*—perhaps the desire to create a good painting

⁸² Mill (1998, Ch. 4).

⁸³ Berridge and Robinson (1998).

⁸⁴ Chalmers (2022, forthcoming).

just because it seems valuable and worthwhile, no matter how it feels. Would the fulfilment of such desires benefit a Vulcan?

The prospect of any improvement in welfare should count in favour of an action. It should give the Vulcan a *justifying reason* for action. It has been argued, however, that a desire without any inbuilt disposition to feel pleasure at the prospect of its fulfilment would fail to provide such a reason.⁸⁵ That is, while affectless desires might motivate behaviour, they would not make the action sensible. Consider for example a Vulcan who suffers from a bizarre condition such that every time it has a radio nearby, it compulsively turns the radio on, without deriving any pleasure or satisfaction from this action or the thought of it.⁸⁶ This functional state—nothing more than a mere disposition to behave—might be called a desire of some sort, but it does not seem to make the action sensible or worth pursuing.

However, while I find this intuitive in the case of turning on radios, cases involving less trivial actions seem to put pressure on the general claim that a desire cannot rationally justify an action unless it also disposes one to feel pleasure at the prospect of its fulfilment. Consider Vulcans' dispassionate desires to perfect their own rational nature, better understand the world, help other beings, or honour their commitments. They may take no pleasure in any of these things, yet believe them to be the right things for them to pursue, and pursue them entirely out of duty. They may believe that this is *what they have to do*, whether they like it or not. When I think about Vulcans in this way, I lose the intuition that they could not act sensibly. As long as they genuinely believe that they are doing these things for their own good or the good of others, it seems sensible to pursue these things. Whether they are mistaken in believing that they can do anything to advance their own good is, of course, the very thing at issue.

⁸⁵ Smithies (2026).

⁸⁶ This is a Vulcanized variant of the *Radio Man* case from Quinn (1993, 237).

Admittedly, I'm not entirely sure whether dispassionate motivational states would qualify as genuine desires, or whether they would rationally justify action. But one thing I'm sure about: affectless desires, if they are desires at all, would not be the paradigmatic kind of desire we encounter in ordinary life. We would therefore risk committing the Extraction Fallacy if we simply took for granted that they retain the welfare significance characteristic of the paradigmatic kind. Instead, we should ask whether it is still intuitively plausible to regard them as relevant to welfare.⁸⁷ I'll make two claims to that end. First, I'll note that, intuitively, only desires about things we care about matter for welfare. Second, I'll argue that, without affect, an entity may act in apparently self-interested ways, but it can neither genuinely care about its goals nor about itself.

Philosophers can use 'caring' however they like, but in the way I use it—which seems to me the most natural—if a subject cares about something, then that thing matters to them, at least to some extent, from their own point of view. Conversely, if something matters to a subject from their own point of view, then they care about it, at least a bit. And if something does not matter to a subject in that way, then they do not care about it (and vice versa). The notion of caring I'm interested in is thus closely related to that of *subjective importance*; I use the terms interchangeably.

For something to be subjectively important to me, as I understand the notion, it is not enough that I believe it is important or take it to be valuable for me. I can believe that making a great sculpture would be something important and valuable for me to do, without caring at all about making one myself or being moved by the idea. Subjective importance essentially involves, among other things, a certain kind of phenomenology directed at a particular

⁸⁷ Chris Heathwood (2019) makes a similar move. After distinguishing between *behavioural desires*, which are intentional states that incline one to act towards a goal, and *genuine-attraction desires*, which involve one's excited attraction towards the goal, he argues that only the latter are relevant to welfare. My argument focuses on the relations among desire, caring, and affect, but it is compatible with Heathwood's if all genuine-attraction desires require caring and affect.

object, such that the object matters to the subject partly in virtue of how it makes them feel (more on this shortly).

This sort of mattering is also bound up with a set of desires about the object: a desire for it to go well, to last, to be present in the subject's life, and so forth. However, we should be careful to distinguish these deep motivational states from shallower ones lacking any genuine subjective importance. The distinction is pervasive in ordinary life and is captured by remarks like 'But I don't really care about it,' made after expressing a preference.⁸⁸

Consider, for example, an art collector who excels at recognising artworks of great aesthetic value. Every week he buys extraordinary new pieces, yet resells them immediately, no matter how beautiful they are. We ask him, 'Why not keep any for yourself?' He replies: 'I think they're good, but I don't feel anything for them. They don't evoke any emotion in me, so I take it that I don't really care about them. It's just my job. And to be clear, I don't care about my job either; I'm simply doing what I'm supposed to be doing.'

We shouldn't be too surprised. Our goal-directed agency and evaluative judgments need not be accompanied by subjective concern. In fact, it seems possible to want something we judge to be valuable, in some detached, impersonal way, even if obtaining it would serve no end we care about. Unusual as this state of mind may be, some people experience something like it in certain forms of depression—not the kind that feels painful, but the kind that quietly makes everything feel empty and pointless.⁸⁹ The art collector might suffer from this condition. He might seem energetic and engaged, always looking for the finest artworks, always judging well, always doing his best—but if he stops for a moment, he can tell that none of it really matters to him. He is just going through the motions.

⁸⁸ Frankfurt (2004, 10–12) draws a similar distinction between things we want or desire and things we care about or regard as important to ourselves.

⁸⁹ Psychiatrists sometimes describe this as *anhedonic depression*—a 'flat' or 'numb' form of major depression characterised by diminished positive affect.

Suppose, further, that the art collector always gets what he wants. Is this good for him? Would he be harmed if someone were to interfere with his aims? Intuitively, it seems to me that he wouldn't. The collector doesn't want the things he wants in a way that puts his welfare at stake. Indeed, it seems to me that if the value of desire-fulfilment were to be understood in such a way that even the fulfilment of goals one does not care about at all could be good for one, then desire-fulfilment would turn out to be a far less plausible candidate welfare good than we ordinarily assume—less plausible, at any rate, than wellbeing theorists or ordinary people would be inclined to accept.

I think that Vulcans are much like this art collector. The only important difference is that, while the collector may have cared about things in the past, may do so again in the future, and presumably remains at least *capable* of caring at present, Vulcans are entirely unable to care about anything in the relevant sense. Caring may require a wide range of cognitive states and dispositions, such that whenever a subject Y cares about ϕ , Y's caring necessarily (i) has some weight in Y's rational decision-making, perhaps enough to warrant significant sacrifices for ϕ ; (ii) draws Y's attention and thoughts to ϕ ; (iii) motivates Y in a robust, enduring way; and (iv) involves a persisting pattern of emotional states and dispositions about ϕ . But what matters for our purposes is that, among all the necessary conditions for caring, one is a disposition to have affective experiences.

Consider:

The Vulcan Mother. A Vulcan mother wants her children to be healthy and goes to great lengths to take care of them. However, she does so without ever having any positive or negative feelings for them, no matter what they do or what happens to them. She is in no way affectively invested in her parental role, never has been, and never will be.

The *Vulcan Mother* coolly exhibits all the typical characteristics of care, without affect. All her thoughts and attention may be focused on her children. She may be disposed to tell herself in inner speech ‘good, that worked’ whenever her children succeed. Her commitment may be unwavering; she might even be willing to sacrifice her own life for them. Yet it would be somewhat odd to say that she *cares about* them. She wouldn’t feel joy at their accomplishments, nor anger at those who intend to harm them, nor grief at their loss. And intuitively, if you care about someone, how their life goes should make you *feel something*. At the very least, it should dispose you to feel something positive or negative at the thought or sight of them faring well or badly. It is for this reason that I think the *Vulcan Mother* wouldn’t be harmed by the loss of her children. Her willingness to make sacrifices for them wouldn’t mean much, because those sacrifices themselves wouldn’t matter to her. And her relentless activity wouldn’t constitute deep commitment to someone she cares about, but rather detached, cold-hearted service.

Some might find it hard to say whether a Vulcan can care about anything in the sense at issue, since it is difficult to imagine the strength and character of the wants of a being whose feelings are necessarily neutral. But there is a way to get close enough. We can start from our own experience and move gradually towards the more alien, Vulcan-like experience. We should pick the closest experience we have to an affectively neutral one: perhaps the experience of staring at a blank wall, something we neither like nor dislike. Keeping this neutral tone fixed inside, we can then go on to imagine entertaining those thoughts and performing those behaviours characteristic of caring. We can think of ways to protect the wall from falling, to paint it anew, and so forth. That is, we can use that familiar unhedonic experience to mentally reconstruct what it would be like to feel no positive or negative affect while still thinking, behaving, and setting goals as we normally do when we care about something. Even if we may never be able to reach, even in imagination, the full extent of affectless agency that might be available to a

Vulcan, we can still imagine being roughly similar in the relevant respects. From this position, we can then ask ourselves: ‘But do I really care about this blank wall—the one I spend so much time thinking about and taking care of, but for which I don’t feel the slightest sadness, grief, fear, joy, or satisfaction, whether it falls or not?’ If we were honest, I suspect we would always answer with a harsh ‘No, not at all.’

Since we wouldn’t care about the wall in those circumstances, we should infer that the *Vulcan Mother* wouldn’t care about her children either. Her case is relevantly like ours. At any time she is with them, she will feel as cold inside as we do when we stare at the wall. And at any time we take care of the wall, we may be exercising to a large extent all the psychological and cognitive faculties she exercises while taking care of her children. Admittedly, she might be better than us in many respects. But it would be strange if a higher degree of dedication, attentiveness, or meticulousness made all the difference to whether or not caring occurs, especially considering that, in the human case, we don’t need to possess those qualities to an especially high degree to care about something. It seems more natural to conclude from this case that the sense of caring at issue necessarily requires some affective investment.

Vulcans may of course reject the human intuitive judgment. They may complain that no one can claim, on their behalf, whether they care about something or not. But humans are in a better epistemic position than Vulcans to evaluate the truth of the necessary link between the relevant sense of caring and affect.⁹⁰ For we can *both* experience affective caring and imagine something close to Vulcan-like caring. By contrast, Vulcans may be inevitably biased. For it may be impossible, or nearly impossible, for them to learn what it is like to care about something affectively.

The above considerations support

⁹⁰ Famously, Mill (1998, Ch. 2) makes a similar argument: in comparing the value of higher and lower pleasures, he takes the judgment of those with experience of both to be epistemically privileged.

The Caring-Affect Link. To care about something in the sense of the term that matters morally and prudentially, one must feel, or be disposed to feel, positive or negative affect at the experience of, or thought of, the object of care faring well or badly.

Put differently, the capacity to feel hedonically engaged with something is *necessary* to care about it in a way that matters. Note that such caring need not involve a disposition to feel affect in response to the object of care *actually* faring well or badly; even just contemplating that prospect may be enough. Consider, for example, a semi-Vulcanized individual who feels nothing when he obtains his goals but feels good whenever he imagines obtaining them. In the relevant sense, he could still care about his goals.

I assume that both hedonic and emotional states can satisfy the Caring-Affect Link, as long as they have some inbuilt felt goodness or badness. Perhaps some emotions are not hedonically charged—perhaps feeling a cold, affectless calm counts as being in an emotional state. But these are not the positive and negative emotional states required for caring, nor the paradigmatic cases of emotion. Indeed, imagining the *Vulcan Mother* feeling glacially calm while her children are dying doesn't make us any more inclined to think that she cares about them or is emotionally attached to them. Vulcans, all in all, are best understood as emotionless beings, since describing them as beings with emotional capacities brings to mind the misleading picture of them experiencing anger, grief, fear, joy, and other paradigmatic emotions that seem to necessarily involve affect.

Let us take stock. I have been arguing that an adequate theory of welfare must discriminate between those desires that contribute to welfare and those that do not, and that only desires about things we care about matter for welfare. Since the sense of caring at issue requires a disposition to feel affect, any being incapable of affective consciousness cannot care about anything in the relevant sense. Thus, Vulcans cannot form desires of the right kind—those

that contribute to welfare. And if nothing matters to a being, so that every choice in its life expresses only cold, emotionally detached purposes, we can reasonably surmise that its aims do not in and of themselves demand moral respect.

I want to close the section with two speculative thoughts about the Caring-Affect Link: one about its nature, the other about its connection to justifying reasons for action.

We might ask: why is there such a necessary link between caring and affect? What grounds it? I have given some reasons to believe in it, but not an explanation of it. I suspect that the deepest explanation has to do with a certain way of representing the value of an object within ourselves. For something to matter to us—for us to care about it—its value must be represented in a certain way in our minds. And to represent value in this way is to register and weigh it in an affective currency. That is, as a matter of metaphysical analysis, subjective importance itself is written in affective characters. I'm not sure how to make the thought more precise.

I also suspect that the Caring-Affect Link throws into doubt whether any desire or intentional state can provide a justifying reason for action without also disposing the subject to feel affect. If you don't care about what you are doing—not even derivatively because you are doing it in order to obtain something else you care about—then why are you doing it? Without caring, it is hard to rationalise your behaviour. It is hard to see why you would have a reason to do anything, why any of your actions would be sensible or worth pursuing. But I'll leave both these thoughts about caring to be developed another time.

4. Other Goods

Let us turn to other candidate welfare goods, such as love, friendship, virtue, knowledge, aesthetic appreciation, achievement, and meaning. Again, we are normally affectively invested in the pursuit of these things—we care about

them and enjoy having them. Perhaps we can have them without affect. But we would risk committing the Extraction Fallacy if we simply assumed their goodness to be preserved. We have stripped away a potentially fundamental component from the familiar conception of these goods. Their intuitive plausibility as welfare goods therefore needs to be tested again. And once it is, I argue, these candidate goods lose all their original appeal. (Similar arguments can be extended to all other possible welfare goods.)

Love and friendship can be treated briefly. Caring is an essential ingredient of these goods, and I have already argued that the relevant kind of caring requires affective investment. Indeed, it is unclear what would be intrinsically good about a Vulcanized form of love or friendship, one deprived of mutual concern, renewed and deepened affection, shared sympathy, amusement, and delight.⁹¹

Virtue is a more interesting case. People with damage to brain areas crucial for affective processing can make sensible moral judgments.⁹² Similarly, a Vulcan could judge things to be good, explain what is good about them, and be motivated to act accordingly. Could such a being be virtuous?

I suspect that virtue requires more than a combination of sound moral judgment and agency. There seems to be nothing virtuous about fighting for a cause you don't care about, or helping a person whose wellbeing doesn't matter to you. More generally, without emotional attachment to sustain patterns of behaviour such as loyalty or costly aiding, it is hard to find them

⁹¹ See the *Vulcan Mother* in Section 3.

⁹² Or at least this is what some scientific studies are often taken to show (Damasio 1994; Bechara et al. 2000). It seems to matter, however, at what developmental stage the brain damage occurs: early lesions to the prefrontal cortex appear to lead to something akin to psychopathy (Anderson et al. 1999). Let me briefly suggest one way to explain this. Perhaps affective consciousness is indeed necessary to learn how to apply moral concepts correctly and to develop the moral sensibilities required for virtue. But when the damage to affective processing occurs after a normal moral development, the subject may still *remember* how to formulate moral judgments and thus behave, at least from the outside, as a normal moral agent would. This might make it seem as though affective consciousness is not required for virtue, when in fact no individual could have acquired the right moral sensibilities without properly exercising their affective faculties at some point in their development.

noble. If genuine interpersonal concern is required for virtue, and the Caring-Affect Link holds, then the virtuous person is partly distinguished by the burning of affective feelings in her.⁹³ For only such feelings can bridge the gap between a merely dispassionate value judgment and a genuine interest in others.

I am open to the Kantian line of criticism that if your primary reason for doing the right thing is grounded in sentiment rather than respect for the moral law, then your act may not be virtuous at all. But even Kantians, I think, should agree that virtuous people need to care about at least some things—such as other people, and perhaps the moral law itself—if they want their view to be intuitively plausible.

Another way to ascribe wellbeing to Vulcans is on the basis of goods such as knowledge. After all, Vulcans can learn many truths about the world and its relations, just as a scientist can understand general relativity without enjoying it. Consider the richest body of knowledge a Vulcan could have:

*Mary the Vulcan Scientist.*⁹⁴ Mary, the most brilliant Vulcan scientist, has dedicated eons to studying the nature of things. Despite lacking affective consciousness, she has supreme knowledge of the machinery of the universe. She knows, for example, exactly which wavelengths from the sky stimulate the human retina, and precisely how these trigger, via the brain, the contraction of vocal cords that produces the exclamation, ‘The sky is beautiful today!’ Is this knowledge good for her?

Mary’s immense knowledge is undeniably awe-inspiring. Yet I think it is greatly limited, in both breadth and depth, by her lack of affective consciousness. Whether she lacks a phenomenal concept, a know-how ability,

⁹³ The Caring-Affect Link thus provides independent support for a sentimentalist conception of virtue (Slote 2009).

⁹⁴ Inspired by Jackson (1982, 130).

or something else, what matters is that she misses a kind of knowledge important to us: she doesn't know what it is like to be in any given affective state. If we asked Mary about Shakespeare's *Romeo and Juliet*, she could offer an impeccable interpretation of the characters' reasons and motives, but she wouldn't know what it is like to love someone. She would be unable to access the affective, qualitative aspects of love from a third-person description. If we asked her about war, she could cite a multitude of historical facts and personal testimonies, but she wouldn't know what it feels like for someone to lose their friends and grieve over the thought that they could have saved them. Even if we turned to science—Mary's crowning achievement—we would find something she doesn't know. Mary herself could never know the thrill of pursuing nature's deepest mysteries, nor experience, in her discoveries, the wondrous understanding that enlightens the mind and stirs the heart. In fact, she could no more imagine what it would be like to feel this way than I could picture my daily life in twenty-seven-dimensional space.

Mary is then ignorant, in my view, of some of the most important things there are to know. Of course, she still knows many other things. But the idea that her remaining knowledge is good for her needs independent support, not an unargued-for statement that knowledge is good. After all, in the human case, coming to know something is usually accompanied by affective feelings. Enlightening intellectual moments are among the most enjoyable experiences a person can have, and we care deeply about having them. Thus when philosophers claim that acquiring knowledge is good for us, what they have in mind may not be knowledge as such, but *knowledge we enjoy or care about*.

This reading is strengthened when we test our intuitions about purely affectless knowledge. It seems odd to claim that if I were to fill your mind with knowledge about the culture of a distant species—knowledge that would never be useful to you, that you would not care about in the slightest, and that would be utterly unenjoyable to you and everyone else—you would nonetheless have reason to be grateful to me. If knowledge is good for the

knower at all, and this thesis is to be intuitive, then all knowledge that is good must involve affective experience.⁹⁵ Mary's knowledge therefore cannot be good for her.

What about aesthetic appreciation? I'll focus on beauty, though similar things can be said about other aesthetic properties. Since Mary has complete scientific knowledge of the human mind and the principles of aesthetic appreciation, she can accurately judge something to be beautiful for humans. She can also recognise when a painting satisfies widely shared norms of beauty. But can she appreciate beauty herself, in the sense that matters for welfare?

I suspect Mary would have to admit, 'I know why people find this painting beautiful, but I just don't see it myself.' For appreciating a work of art as beautiful requires more than knowing that it is beautiful. It requires experiencing the piece with affective feelings directed at its aesthetic properties, such that the experience *feels heavenly* in the distinctive way appreciating beauty *feels heavenly*. Mary could not learn what this experience feels like via introspection, and it is notoriously difficult to conceive how she could derive this first-personal knowledge from the rest of her knowledge. Were she to suddenly gain affective consciousness, she would appreciate beauty for the first time.

Let us turn to the value of achievement. Like human athletes, Vulcans can train relentlessly, master their skills to perfection, and order their lives around the pursuit of excellence. Our conception of achievement, however, is normally associated with feelings of deep satisfaction, fulfilment, or at least excitement along the way. Strip all that away. When Vulcans work towards achieving something, they don't enjoy the process, they don't feel satisfaction when they succeed, and they don't care about the overarching goal at all. To me—and, I suspect, to most people—it is doubtful that such an achievement could be intrinsically beneficial to the achiever.

⁹⁵ A view of this kind is examined by Kagan (2009).

Now, one might think that what makes an achievement good has nothing to do with the feelings it brings or how much the achiever cares about it, and everything to do with the value of perfecting the achiever's own nature. There is some charm to this idea: through superlative achievements, a creature can purposefully develop its capacities to the fullest, making its life excellent and fulfilling its *telos*.⁹⁶

Presumably, however, perfecting a given capacity is not always good for the one who has it. A capacity is worth perfecting for the sake of its possessor only if it has at least the *potential* to be good for them. Perfecting the capacities of a non-conscious robot, for instance, cannot be good for the robot itself, because those capacities have no such potential. So for a Vulcan to benefit from perfecting its capacities, one would first need to show that a conscious being who lacks affective faculties can still have capacities that are potentially good for it. But the Vulcan wouldn't enjoy its capacities, appreciate their beauty, or care about them at all. There seem to be no intuitive grounds for thinking that such capacities have any potential for prudential goodness. And it cannot be good for the Vulcan to perfect something that can be, at best, neutral.

The last candidate welfare good I wish to consider is meaning. It seems good for one to have a meaningful life, and there is some reason to suspect that a Vulcan could have one. After all, meaning seems largely independent of how well one is doing from a purely hedonic point of view. Viktor Frankl, for example, maintained that some concentration camp inmates preserved a sense of meaning even in the midst of extreme suffering.⁹⁷ Other examples of meaningful suffering include caring for an elderly relative at the expense of one's own enjoyment, or struggling against injustice at great personal cost. If suffering does not preclude meaning, why should mere neutrality? If even lives of suffering can be meaningful, a merely neutral life like a Vulcan's might seem

⁹⁶ Bradford (2015).

⁹⁷ Frankl (1959).

all the more capable of achieving meaning through non-hedonic purposiveness.

We could imagine a *Vulcan Tolstoy*, identical to the Russian author in every way except for lacking his existential despair and any other positive or negative affective state. Though it writes books just as well and changes the history of Vulcan literature just as much, there are good grounds to believe that only the human Tolstoy had a meaningful life. Part of what imbued Tolstoy's life with meaning was how hard he struggled to understand the human condition and how one ought to live *in spite of* his torment and despair. He was deeply drawn to this pursuit, so deeply that he was willing to pursue it no matter how much he was suffering. It was just *that* important to him. I suspect that this deep subjective attraction to a goal is an inescapable component of what makes it meaningful. For if I were to discover that Tolstoy had dedicated his life to a body of work that didn't matter much to him, that he spent his days without caring particularly about what he was doing, and that he would have done something else had he faced the mildest suffering, I would immediately cease to see his life as one of meaning.

In other words, I am suggesting that to have a meaningful life, a person needs to be deeply affectively engaged with their pursuit. They must be engaged in such a way that they would strive to pursue it even at the cost of going through a lot of suffering or other great personal costs, because the goal *matters* to them that deeply. Perhaps the object of pursuit must also be objectively important, from the point of view of the universe, so to speak. Indeed, it has been argued that meaning involves attachment to worthy objects: doing something objectively valuable in a way that is subjectively attractive.⁹⁸ But whether or not this objective condition holds, the subjective condition is enough to rule out that Vulcans could have meaning. If a meaningful life is a life absorbed, stimulated, or consumed by a certain goal—say, producing art or helping others—in a way that matters deeply to the one

⁹⁸ Wolf (2010).

who pursues it, then only affective beings can have meaning. For only a being who deeply cares about something can be attached to it in a meaning-conferring way; and only a being disposed to feel deep positive or negative emotions at the experience or thought of the object of pursuit can be said to care about it deeply.

5. Objections

I have argued that every promising candidate welfare good or bad has an affective component, and therefore that only affective beings can be welfare subjects and have moral status. I'll now consider five objections: that becoming a Vulcan seems better than death (5.1); that Vulcanization seems the goal of certain contemplative traditions, such as Buddhism (5.2); that killing Vulcans seems wrong (5.3); that autonomy can ground moral status even in the absence of welfare (5.4); and that having a neutral level of welfare suffices for being a welfare subject (5.5).

5.1 Vulcanization and Death

If I had to lose, irreversibly, either my entire phenomenology or just its affective properties, which would I rather lose? Suppose everything else is equal. From the outside, I would keep behaving as I always do. But on the inside, I must choose: my whole consciousness or my affective life?

If affective consciousness is necessary for being a welfare subject, I should be indifferent between the outcomes. I wouldn't get anything good out of life either way. Indeed, I should be indifferent between these options and death, at least as long as I know that, after I'm gone, a robot will fulfil my plans and take care of my loved ones as well as I would have done. For some, this will be hard to believe.⁹⁹ And, not unreasonably, they could take their intuitive response at face value—as evidence that consciousness itself is a welfare good,

⁹⁹ Kriegel (forthcoming).

or at least enables one to be a welfare subject for whom things other than consciousness can be intrinsically good.¹⁰⁰

There is also, however, a way to account for this intuition that is entirely compatible with the idea that only affective beings are welfare subjects. I might be attracted to a metaphysics of persons on which I can continue to exist after the irreversible loss of my affective faculties, whereas I would cease to exist after the irreversible loss of my very own conscious point of view. And I may care a lot that it remains ‘me’ who carries out my plans, fulfils my current preferences, and stands in certain relationships to my loved ones. Choosing Vulcanization would be, in some ways, relevantly like fulfilling a particular preference concerning how I want to die. Even if I know that, whether through Vulcanization or death, I will cease to be a welfare subject, I could still have a reason to die as I want, or to be Vulcanized if I want. For more of the desires I have while I am still a welfare subject would be fulfilled in this way, thereby giving me a higher level of wellbeing at the time I am a welfare subject.

Admittedly, this way of addressing the challenge doesn’t fully satisfy me. It relies not only on a particular view of the metaphysics of persons and the constituents of welfare, but also on the presence of certain desires that seem to me unjustified, if not utterly irrational. Why would it be so important to me that I am the one carrying out my plans and fulfilling my preferences, rather than my non-conscious body or a non-conscious robot, when I would be only an empty shell of what I used to be—a piece of meat that would in a strict sense still be me, but would lack everything that made my own existence valuable and worthwhile and important to me?

I prefer a different diagnosis of the intuition that we would rather be Vulcanized than lose consciousness irreversibly. It rests, I think, on a failure to imagine sufficiently well what it would be like to be Vulcanized. This is understandable. People report having positive or negative feelings or moods at

¹⁰⁰ Against the view that consciousness itself is a welfare good, see Lee (2019).

almost any moment, so it is difficult for us to imagine life without them.¹⁰¹ As I suggested earlier, the best way to imagine such a life is to examine your own phenomenology and pick an experience that feels as neutral as possible, one with negligible or non-existent affect. For me, this is staring at a blank wall. You can then imagine going on doing things—perceiving colours, forming thoughts, and moving your body—while retaining the same affective tone on the inside, such that even during what are supposed to be the most important moments of your life, you feel as you would feel staring at a blank wall.

When I imagine a life like this, I think I might still be doing some things. I might remember my past because I need to, take care of my loved ones because I have to, and write about philosophy because I'm used to doing so. But I would no longer care about any of these things.¹⁰² For whatever happens, however wonderful or awful, it would not produce in me the smallest flicker of happiness or sadness, excitement or distress. My world would be cold, emotionally detached from everything and everyone. It's not hard to believe, at least for me, that I would gain nothing from my days, and that the thought of continuing to exist would itself lose its appeal.

Losing affective consciousness, irreversibly and in all its forms, would then be prudentially relevantly like dying. It would amount to the loss of my status as a welfare subject and to the deprivation of all the goods and bads I would otherwise have accrued had I remained one.¹⁰³ This does not mean that

¹⁰¹ Lucas and Diener (2008, 796).

¹⁰² Kriegel (forthcoming) considers the possibility that without affect nothing would matter to us, including our own survival. But he rules out this hypothesis on the grounds that we can be said to care about something without feeling good or bad at that very moment. I agree that caring need not involve occurrent feeling at every moment, but I hope to have shown in my earlier discussion of the Caring-Affect Link that, whenever we are said to care about something, we must at least be *disposed* to feel good or bad in response to whether the object of care fares well or badly.

¹⁰³ One might think that dying is *intrinsically* bad in a way that the mere loss of affect is not: I would cease to exist, and my whole consciousness would be annihilated, along with my character, intentions, and beliefs. But this intuition too, in my view, rests on a failure to imagine clearly enough what it would be like to be Vulcanized. For the annihilation of all my enjoyments, all my cherished emotions and moods, and my very capacity to care about anything—including my

people could treat me however they want, just as they cannot treat a dead body however they want. It is still possible, for example, that what happens to me could affect how many of the desires I had while I was a welfare subject are fulfilled, since presumably I had some desires about how I and my body should be treated. If so, people could harm a Vulcanized individual in the same sense in which they could inflict a posthumous harm—by frustrating the desires of the original welfare subject, thereby lowering that subject’s wellbeing level and harming them at a time when they could still bear welfare goods and bads.¹⁰⁴ Those who reject the possibility of posthumous harms have other ways to explain why certain ways of treating a Vulcanized individual are wrong. They could appeal to our obligations to respect certain norms or cultural values, to honour the wishes of those who still care about that individual, or to cultivate certain virtuous character traits. On none of these explanations does the Vulcanized individual need to have any interests of their own.

5.2 Vulcanization and Buddhism

Some might see an analogy between the affectless mental states characteristic of a Vulcan’s life and those pursued in the Buddhist tradition. The goal of Buddhist practice, after all, is to free oneself from the dissatisfaction (*duḥkha*) that arises from attachment (*trṣṇā*) to what one takes to be good for oneself and aversion (*dveṣa*) to what one takes to be bad for oneself. I have argued that if nothing can feel good or bad to a given subject, nothing can be of subjective importance to them—this is the core idea behind the Caring-Affect Link. Becoming a Vulcan, were this possible through some procedure, would then seem to be an especially effective way to free oneself from any substantive form of attachment or aversion. Seen in this light, the Vulcan’s hedonically neutral way of experiencing life would turn out to be one of the best possible states of

character, intentions, and beliefs—might seem just as intrinsically bad. Indeed, it might explain why dying seems intrinsically bad at all, since to lose consciousness is to lose affect.

¹⁰⁴ For a more extensive discussion of posthumous harms, and of whether they are compatible with the idea that only conscious beings have interests, see Chapter 3.

mind, according to nothing less than one of the most ancient and widely practised philosophical traditions.

I honestly find it difficult to see how the goal of any tradition in moral philosophy could be the severe, complete removal of the practitioner's affective dimension of experience, unless one assumes a form of pessimism about the value of life on which life is mostly bad and we would all be better off if we were simply incapable of accruing any positive or negative welfare at all.¹⁰⁵ This pessimism, importantly, would be entirely compatible with the idea that only affective beings can be bearers of welfare goods or bads. It would imply only that such beings are worse off in a life of suffering than in a merely neutral one.

Perhaps there is such a pessimistic assumption at the basis of the Buddhist view. But a closer look at its traditional texts seems to suggest something different: that there is something emphatically positive about the state of mind to which we should all aspire, and that this state does involve affective experience—just a form of affective experience different in some ways from the most ordinary kind.¹⁰⁶ In the *Māgaṇḍiya Sutta*, nirvana is referred to by the Buddha as the highest happiness and greatest bliss (*paramaṃ sukhaṃ*). In Abhidharma philosophy—which attempts to give a systematic reconstruction of what the Buddha taught in the discourses—some kind of hedonic tone (*vedanā*) is said to be a part of every mental episode, even for awakened beings, those who have attained enlightenment. So, at least on one plausible interpretation, the goal of Buddhist practice is not to free ourselves from affective states altogether. It is to reach a certain wholesome, positive quality of experience, often described as full of loving-kindness, unsurpassed tranquillity, and compassion. And the only way to reach this state is to free ourselves from a certain inappropriate exercise of our affective faculties—an inappropriate intentional attitude towards the contents of

¹⁰⁵ Against this pessimistic reading of Buddhism, see Baker (2024).

¹⁰⁶ On the Buddhist conception of wellbeing, see Baker (2023).

affective experience that makes us crave things, or be averse to them, in ways that are overall bad for us. This phenomenological reorientation never requires us to abandon the affective dimension of our lives; rather, it is supposed to allow us to get the best out of it, in accordance with the Buddhist understanding of what is valuable.

5.3 Killing Vulcans

If I could kill a Vulcan for a small personal gain, doing so might seem morally wrong. It might seem wrong even if there were no instrumental reasons to keep the Vulcan alive, as plenty of others could take its place. On this basis, some philosophers are inclined to conclude that Vulcans have moral status.¹⁰⁷

Admittedly, I once shared the intuition that it would be wrong to kill a Vulcan. But after reflecting more carefully on what it is like to be one, that intuition has vanished. I now see that there would be nothing good for a Vulcan in continuing to exist. My earlier intuition was a result of my imagining Vulcans as more like us than, in fact, they are.

If your intuition hasn't vanished, you might reconsider whether it really tracks moral status. Perhaps killing a Vulcan seems wrong for the same reason that destroying a beautiful artwork seems wrong: some things might have impersonal value even without moral status.¹⁰⁸ Or perhaps we should expect people to be reluctant to kill Vulcans in much the same way that research participants are reluctant to drink Mountain Dew from a bedpan. We habitually associate complex intentionality with affective consciousness, so our intuitions may well misfire when confronted with alien affectless minds. Finally, even if Affective Sentientism is true—and you may be largely persuaded that it is—the arguments I've given may not make you *certain* of it.

¹⁰⁷ Chalmers (2022, 335–9), Birch (2022, 799), Mogensen (2025b).

¹⁰⁸ Smithies (2026).

And under significant normative uncertainty, one ought not to kill a Vulcan for a trivial end.¹⁰⁹

5.4 Autonomy and Caring

So far, I have assumed that only welfare subjects have interests worth moral consideration. On some views, however, beyond an interest in welfare, we also have an interest in *autonomy*—in guiding our lives according to our own will and rational reflection. It often seems wrong, for example, to intervene paternalistically in ways that would foreseeably improve someone’s welfare but disregard their autonomy.¹¹⁰ Once we allow that some interests are independent of welfare, it stands to reason that the properties responsible for those interests might be instantiated in an individual even when the properties that make that individual a welfare subject are absent. Some autonomous agents, then, might have moral status without being welfare subjects.

Vulcans could be an example of such beings.¹¹¹ Although they cannot be harmed or benefited, perhaps they can still be wronged by acts that interfere with their rational self-governance; hence killing one would be wrong.

Let us grant that Vulcans might be autonomous agents. I nonetheless see a problem: not every form of autonomy seems worthy of respect. To my mind, an agent’s autonomy deserves respect only when the agent *cares about* the goals they set for themselves.

Consider, for example, asking a friend whether he would like to order a pizza margherita or a pizza carbonara. He replies, unexcitedly, ‘This choice doesn’t matter to me in the slightest.’ You know he isn’t joking. He still picks the carbonara—partly out of habit, partly because he mildly prefers the taste—but insists he wouldn’t be the least bit bothered if he ended up with a margherita. ‘In fact, I wouldn’t even care if you ignored my preference,’ he

¹⁰⁹ For a probabilistic precautionary approach to cases of uncertain moral status, see Sebo (2025).

¹¹⁰ Enoch (2016).

¹¹¹ This is Mogensen’s view (2025b).

adds. To his surprise, this is precisely what happens. You order margheritas for both because doing so makes it cheaper for whoever ends up paying. You explain to your friend that you didn't mean to disrespect his general autonomy, which certainly deserves consideration on other occasions. You simply noticed that, on this occasion, he didn't care at all. Have you wronged your friend? It doesn't seem so. Your interests can easily override such shallow preferences. Generalising from this case, it seems that there is no moral reason to respect others' autonomy except insofar as their choices express preferences they care about.

Of course, humans usually care about their preferences. But I have shown that only affective beings can do so in the relevant sense. Only such beings, then, can exercise their autonomy in a way that deserves respect. And since all affective beings are welfare subjects, it follows that only welfare subjects have moral status—not because all morally relevant interests are grounded in welfare, but because affective consciousness is necessary for both welfare-based and autonomy-based interests. This also strikes me as an intuitively compelling result in its own right: I don't see anything morally wrong in overriding the self-given ends of a creature to which nothing feels good, valuable, or important.

5.5 Neutral Welfare and Welfare Subjects

I have argued that nothing can be good or bad for Vulcans, and that they are not welfare subjects for this reason. Even so, it still makes sense to ask how desirable it would be to be one. We can imagine taking up a Vulcan's conscious perspective in order to evaluate whether such beings are better or worse off than we are. If such comparisons are possible, then Vulcans must have a level of welfare: a neutral or zero level. For we could not be better or worse off relative to beings whose welfare level is totally unspecified. But now it seems we have a problem. What if to have a welfare level just is to be a welfare

subject?¹¹² Then Vulcans would turn out to be welfare subjects after all. They would in fact have a welfare level of zero rather than none.

I think we should resist this conclusion. If even entities whose welfare level can only be zero count as welfare subjects, it becomes hard to see what work the notion of welfare subject does in ethics. There is nothing at stake, morally, in the case of entities that cannot have welfare goods or bads but can have a neutral level of welfare. They are not morally different from entities that lack welfare altogether. After all, there is no important difference between a being who can hear only silence and one who cannot hear at all—either way, they miss out on the music.

Because I care about maintaining a robust link between welfare subjectivity and normative ethics, I restrict the notion to those entities that can have welfare goods or bads. Of course, one might also go the other way and treat welfare subjectivity as a non-moral property that some beings possess despite lacking moral status. But on this view there would still be a subset of welfare subjects—the *morally significant* welfare subjects—who do have moral status. My arguments could then be taken to support the thesis that these morally significant welfare subjects are all and only affective beings.

Now, it is not obvious that beings whose welfare level is fixed at zero do not matter morally. They might matter in virtue of their beauty, their usefulness, or other contingent features. What I want to deny is that they matter *in virtue of their having a neutral level of welfare*. Let me defend this claim against some challenges.

First, consider *Net-Zero*: a being whose welfare goods and bads happen to be perfectly balanced at every moment, so that its overall welfare level is always neutral.¹¹³ One might take *Net-Zero* to demonstrate that having a neutral welfare level suffices for having morally relevant interests. But *Net-Zero* is a welfare subject because it can—and in fact does—have welfare goods

¹¹² van der Deijl (2021, 2025), Lee (2025), Kriegel (forthcoming).

¹¹³ For similar cases, see Lee (2025, 13) on *Balanced* and Kriegel (forthcoming) on *Mombies*.

and bads. It is that capacity, not the neutral welfare level itself, that grounds its morally relevant interests and specifies their content.

Second, consider a Vulcan-like robot, conscious but incapable of positive or negative affect. Call it *Anhedonica*. Now, if we had a pill that could endow any being with the gift of positive affective experiences, wouldn't we have a strong moral reason to give it to *Anhedonica* rather than to a non-conscious robot?¹¹⁴ And wouldn't that reason be grounded in the fact that *Anhedonica* has a strong interest in receiving the pill, given its neutral—and therefore low—level of welfare?

I see two ways of addressing this case without abandoning the idea that only affective beings have interests. One possibility is that a being can already be said to be *capable* of affect, at least in the morally relevant sense, if the conditions are such that it can easily acquire affective states. *Anhedonica* would thus be a welfare subject *before* receiving the pill and, as such, would have an interest in enjoying the experiences that the pill would provide. The non-conscious robot, by contrast, would seem to lack any such interest. For by giving it the pill, and thereby creating a new locus of consciousness, we would seem to be bringing an entirely new individual into existence (though this result will depend on the exact relation between consciousness and personal identity).

That *Anhedonica* may already be capable of affect shouldn't surprise us. Our capacities can and often do depend on external facts. We wouldn't be able to do many of the things we are said to be able to do without oxygen or gravity, and yet these are not intrinsic properties of human nature. The pill, if easily accessible, could be just like oxygen: as long as it is around, and perhaps only then, a being can be said to have certain capacities.

There is also another way to address this case. Suppose that *Anhedonica* turns out to lack the capacity for affect in any morally relevant sense of

¹¹⁴ van der Deijl (2025).

‘capacity.’ In that case, I think we should simply deny that *Anhedonica* has interests. We can do so on the basis of

The Sake Condition. An entity has interests just in case it has a *sake*—that is, just in case it has a perspective from which things can be good or bad for the entity itself.

Only welfare subjects have a sake, and so only welfare subjects have interests. To be clear, receiving the pill would indeed benefit *Anhedonica*: post-pill *Anhedonica* would have higher welfare than pre-pill *Anhedonica*, and there is continuity in consciousness, so the very same individual would be made better off. What I am claiming is that pre-pill *Anhedonica* does not have an *interest* in being benefited. Even though ordinarily whenever someone is benefited they have an antecedent interest in the benefit, the two can in principle come apart. An entity can be benefited by the very event that creates its sake, and in such a case it would not have had an antecedent interest in the benefit. For one cannot have anything at stake, and thus any interest, if one does not even have a perspective from which things can be good or bad for one. *Anhedonica* will acquire this perspective only when, and if, it receives the pill. Until then, just as a rock is perspectiveless with respect to having a conscious point of view, *Anhedonica* is perspectiveless with respect to having a sake. It has a conscious perspective, to be sure—but not the kind that would make it a bearer of interests.

Of course, there is an obvious sense in which things *can* be good or bad for *Anhedonica*, given that it *can* receive the pill. But this is the same sense in which we could say that it *can* have affective experiences. Thus, if we take this to be the morally relevant sense of ‘capacity,’ pre-pill *Anhedonica* would count as both a welfare subject and an affective being, as noted earlier. Here, however, we are assuming that the relevant notion of capacity works in such a way that *Anhedonica* should be said to lack the capacity for affect even though it can receive the pill. Absent any reason for a double standard, we should

similarly say that *Anhedonica* lacks the capacity for welfare goods or bads in the morally relevant sense of ‘capacity,’ even though it can receive the pill in the mere sense of modal possibility. It follows that pre-pill *Anhedonica* lacks a sake: it lacks a perspective from which things can be good or bad for it in the relevant sense of ‘can.’

Who should then get the pill—*Anhedonica* or the non-conscious robot? Neither of them has any interest in being benefited before receiving the pill; they both lack the faculties necessary to have a sake. So in neither case can we have a moral reason to benefit them grounded in their currently existing interests. Any such reason, then, must be grounded in their future interests—in the fact that they will be grateful for what we did and what they gained from it. But this applies equally to both beings. Both *Anhedonica* and the robot will be every bit as glad for what we did. Therefore, all else equal, I think we should be indifferent as to which of the two receives the pill.

It is controversial to claim that there is no difference between creating a happy being and increasing the happiness of an already existing being. But this is a special case, because the already existing being is not a welfare subject—it exists, but not as a subject that matters morally. So, for moral purposes, giving the pill to *Anhedonica* is relevantly like bringing a new being into existence: we are creating a sake and a life worth living where there were none before, and bringing about interests where there were none before.¹¹⁵

One might suggest that pre-pill *Anhedonica* has some derivative interests in virtue of the fact that it *would have* interests were it to receive the

¹¹⁵ The moral significance of the distinction between making people happy and making happy people is often thought to turn on the fact that in only the former sort of case does failure leave someone with a complaint. On this basis, someone might object that there is a difference between giving the pill to *Anhedonica* and bringing a new happy being into existence: crucially, if we fail to give *Anhedonica* the pill, *Anhedonica* will be in a position to complain that it is worse off than it could have been, whereas if we fail to bring someone into existence, there is no one in a position to make that sort of complaint. My view, however, is that only beings that have a perspective from which things can go better or worse for them can complain, in a morally relevant sense, that they are worse off than they could have been. Pre-pill *Anhedonica* does not have a sake, and so could not make any morally relevant complaint were it not to receive the pill.

pill. But the idea of attributing interests on the basis of mere modal potential would have implausible implications when applied to human fetuses or non-human animals such as insects, which could also gain human-like forms of sentience, rationality, autonomy, and the corresponding interests with the right pill.¹¹⁶

One last challenge to the view that beings whose welfare level is fixed at zero do not matter morally. When we calculate the average welfare in a world, it could be argued that we must factor in such beings. This would have implications for non-separable population axiologies, such as averagingism.¹¹⁷ However, they would be *odd* implications. In a terribly unhappy world, for example, I might have a strong moral reason to create many Vulcan-like AIs to improve the average. But setting aside any indirect benefits, I cannot bring myself to believe that I would be ‘making the world a better place’ by bringing into existence beings that cannot even have lives worth living. Worse still, on these views I could improve the world by creating beings that can only suffer, so long as the existing population suffers even more. I suspect that any population axiology that gives significant weight to Vulcan-like beings has wildly implausible implications—worse than those of competing views—and should be rejected.

6. Conclusion and Implications

I am conscious, therefore I matter? Attractive as the idea may sound, it is mistaken. Moral status requires more than mere consciousness: it requires consciousness with a positive or negative hedonic valence. And even if beings like Vulcans do not yet exist, this conclusion has real implications.

For one thing, whenever we try to establish the moral status of beings different from us, we now need to determine whether they have affective states, not merely whether they are conscious. In part, this makes things easier:

¹¹⁶ McMahan (2016).

¹¹⁷ Lee (2025, 12).

it shifts our attention to well-studied targets of empirical investigation such as emotion, pain, and pleasure. In part, it makes things harder: we must ascertain the presence of *both* consciousness *and* affect.

Does this make things too intractable for a hugely consequential concept like moral status? Given how poorly we understand consciousness, it has been suggested that the investigation of cognition-generated interests is all that should be required.¹¹⁸ But I take my earlier arguments to have isolated what really matters: affective consciousness. Without it, cognition isn't enough for moral status. And although this complicates attributions of moral status, a significant probability of affective consciousness already calls for caution, and assigning such probabilities seems empirically tractable.

We can already tell that all vertebrate animals are likely to be affectively conscious. They exhibit strong indicators of consciousness, and affect is among the most ancient evolutionary features of the mind. We remain uncertain whether invertebrates are conscious, but it seems plausible that if they meet the conditions for consciousness, they also meet those for affect. On Affective Sentientism, then, all vertebrates—and perhaps many invertebrates—have moral status.¹¹⁹

Things are different for artificial intelligence systems, with which we share no evolutionary history. Such systems may soon exhibit cognitive and sensory processing, a global workspace, and unified agency. With these capacities, AIs could well turn out to be conscious, goal-directed rational agents. But it is far less clear whether these capacities entail affective consciousness. We could therefore have conscious AIs without affect—AIs that, on Affective Sentientism, would lack moral status.

The possibility of Artificial Vulcans might shape the future of AI design. These artificial systems could reason, imagine, plan, understand, act,

¹¹⁸ Levy (2020).

¹¹⁹ On the evidence for affective consciousness in vertebrates and invertebrates, see Birch (2021, 2024).

and govern themselves. Yet they could not suffer or be wronged. We could delete, copy, or replace them without worrying about their interests. This is not to say that we may treat them however we like, or that shutting them down is always permissible. But any reasons against doing so must be grounded in indirect effects on affective beings, institutions, or norms—not in harms to the artificial systems themselves. Artificial Vulcans thus represent an opportunity to make use of the benefits of advanced intelligence and consciousness without the dangers of suffering and exploitation.

The benefits of AI consciousness may be considerable. There are many tasks that an AI may not be able to perform without being conscious, or at least that might prove harder to engineer without it. At a sufficiently high level of sophistication, an AI's capacity to integrate multimodal information, monitor its own processing, and coordinate its internal states and goal-directed behaviours under a unified perspective may provide strong evidence of consciousness, if not suffice for it outright. If AI consciousness turns out to be unavoidable or desirable, we will need to know whether it can be developed without thereby creating new welfare subjects—and, if so, how. It may then be worth pursuing more research into non-hedonic motivational architectures, so that we might one day design conscious AI systems without affective processes, and do so safely.

Chapter 3:

On What Matters to the Dead

Can a person ever be harmed by events occurring after her death? This chapter puts forward a new theory of posthumous harm, which builds on Derek Parfit's distinction between identity and what matters in survival. Whatever our views are on honouring the dead, post-mortem interests, and the desires of the past, we all agree that certain things matter in survival: there is a relation that underpins the special, prudential concern we ordinarily believe we should have for ourselves. I argue that if we also agree with Parfit that these things matter even when identity is not preserved, we can explain the possibility of posthumous harms and benefits just by appealing to the nature of prudence. Although we might no longer be in a position to harm individuals identical to a dead person, we can harm others who are suitably psychologically related to her—the dead person's continuers.

1. Introduction

A dead man is popularly believed to be capable of experiencing both good and ill fortune—honour and dishonour, and prosperity and the loss of it among his children and descendants generally—in exactly the same way as if he were alive but unaware or unobservant of what is happening.

Aristotle, *Nicomachean Ethics*, 1.10

Before dying, many strive to pass on the things they care about to others. Aristotle, for instance, devoted his life to the creation and transmission of an immense body of philosophical work. Diogenes Laërtius reported that it comprised 400 books totalling 445,270 lines of writing. Sadly for Aristotle, however, all the books intended for the public were lost or destroyed after his death. The survival of Aristotelian teachings was only possible thanks to the relentless copying-work of generations of monastic scribes, who preserved some of Aristotle's most technical manuscripts throughout the Middle Ages. But suppose things had gone differently. Suppose that those scribes had regarded the remaining writings of Aristotle as unacceptably dangerous and decided to burn them instead. This would arguably have made things worse for those who would have benefited from reading Aristotle, and for those who would have indirectly benefited from the overall advancement of philosophy—perhaps of humankind—due to his influence. But could the destruction of his life's work also make things worse for the deceased Aristotle? If, as I shall assume, death amounts to the end of us, the permanent extinction

of our consciousness, could Aristotle be harmed by events taking place centuries after his passing?

Many philosophers think so.¹²⁰ Death ended Aristotle's existence but didn't end his susceptibility to harm. There can be harms that occur after a person's demise—*posthumous harms*—as well as *posthumous benefits*. How this is possible, though, is deeply puzzling. If Aristotle no longer exists at the time of a posthumous harm, how can he be the one who is harmed? A suitable subject of harm seems missing. I call this the *No-Subject Problem*.¹²¹ It is a conjunction of three claims:

Identity-Tracking Harm. In cases of posthumous harm, if there is any, the deceased is the one who is harmed.

Existence Requirement. An individual can be harmed only if they exist at the time of the harm.¹²²

Dead Non-Existence. The deceased does not exist at the time of any alleged posthumous harm.

One way of dealing with the No-Subject Problem is to reject the possibility of *posthumously timed harms* (that is, harms that occur to a subject at posthumous times, negatively affecting his wellbeing level at such times). Post-mortem persons are, after all, little more than dust, and dust cannot be harmed. Instead, posthumous harms could be understood as negative effects on a subject's wellbeing that occur at a time when the subject still exists, like

¹²⁰ Among them are Nagel (1970), Feinberg (1974, 1977, 1984), Brandt (1979), Pitcher (1984), Parfit (1984), Levenbook (1984), Griffin (1986), McMahan (1988), Grover (1989), Serafini (1990), Luper (2004), Sperling (2008), Belliotti (2011), Nussbaum (2013), Boonin (2019), and Aristotle himself (NE 1.10).

¹²¹ Feinberg (1984, 80).

¹²² McMahan named it the Existence Requirement (1988, 33), and Feldman called it the Existence Condition (1991, 205).

any other harm. The best-known theory of posthumous harm—the *Desire-Fulfilment View*—does exactly that, rejecting *Dead Non-Existence*.¹²³

The Desire-Fulfilment View is derived from a theory of wellbeing that includes desire-frustration as one of its *welfare bads* (what is intrinsically bad for an individual) and desire-fulfilment as one of its *welfare goods* (what is intrinsically good for an individual).¹²⁴ According to this view, if event X frustrates the desire of individual Y, then X harms Y, all else equal. This applies even if X occurs after Y's death, since X would still make it the case that Y had an unfulfilled desire while alive. The subject of posthumous harm is then the formerly living person, whose ante-mortem wellbeing level is made lower than it would otherwise have been.¹²⁵ What makes these harms *posthumous* at all is solely their dependence on posthumous events. Consider again the example of Aristotle: if how well his life went at least partly depends on his desires being fulfilled, and this fulfilment partly depends on things occurring after his death, then it follows that how well Aristotle's life went partly depends on things occurring after his death.

However, deriving a theory of posthumous harm from a particular theory of wellbeing makes it vulnerable to serious objections. There are major

¹²³ Major proponents of this view are Pitcher (1984), Feinberg (1984), and Boonin (2019).

¹²⁴ Note that proponents of the Desire-Fulfilment View of wellbeing can resist the possibility of posthumous harms in various ways. For example, they may endorse an *awareness requirement*, according to which a subject must be aware of an event for it to benefit or harm them, or accept *concurrentism*, the requirement that a subject must hold the desire at the time it is either fulfilled or frustrated for that fulfilment or frustration to benefit or harm them.

¹²⁵ Typically, as a theory of welfare, the Desire-Fulfilment View isn't assumed to come equipped with any inbuilt answer to the question of when harms involving desire-frustration occur. Instead, there are a range of options: the time when the desire is held, the time when the undesired state of affairs obtains, or something more complicated. But in the debate on posthumous harm, philosophers have primarily understood the harm of desire-frustration as occurring when the subject is still alive. On Pitcher's view (1984), the subject suffers the posthumous harm before death, more or less at the time at which she forms the desire that will later remain unfulfilled. Boonin (2019) also believes that the subject of harm is the living subject (in contrast to either the individual's lifeless remains or a property of the individual that survives after death), but his view is compatible with the absence of an exact time at which, as a consequence of a post-mortem event, the living subject suffers a harm.

problems with the Desire-Fulfilment View of wellbeing in general,¹²⁶ and how it applies to posthumous harm in particular.¹²⁷ Even granting that this theory does not obviously imply physical backward causation,¹²⁸ it still implies an odd backward determination of welfare value, one made particularly odd by the remoteness of the events that could retrospectively influence this form of personal value: how well my own life is going for me *right now* could hinge on states of affairs occurring centuries and centuries after my death.

This doesn't mean that posthumous harms are pure and simple fantasy. It could just mean that the right account hasn't yet been found. This chapter puts forward a new theory of posthumous harm, which builds on Derek Parfit's distinction between *identity* and *what matters in survival*.¹²⁹ Whatever our views are on honouring the dead, post-mortem interests, and the desires of the past, we all agree that certain things matter in survival: there is a relation that underpins the special, prudential concern we ordinarily believe we should have for ourselves. I argue that if we also agree with Parfit that these things matter even when identity is not preserved, we can explain the possibility of posthumous harms and benefits just by appealing to the nature of prudence. Although we might no longer be in a position to harm individuals identical to a dead person, we can harm others who are suitably psychologically related to her—the dead person's *continuers*.

Parfit himself appreciated the implications of his view for posthumous interests in a famous passage:

¹²⁶ Arneson (1999), Crisp (2001).

¹²⁷ Portmore (2007), Jenkins (2011).

¹²⁸ Advocates of the Desire-Fulfilment View draw a distinction between an act resulting in a physical change in the world and an act being responsible for making a particular proposition true. The post-mortem event simply makes it true that at an earlier time the living subject was suffering a harm.

¹²⁹ Serafini (1990) already noted how a Parfitian metaphysics of persons might offer a fresh understanding of posthumous harms. Here I intend to develop and articulate this theory in detail.

When I believed that my existence was such a further fact, I seemed imprisoned in myself. My life seemed like a glass tunnel, through which I was moving faster every year, and at the end of which there was darkness. When I changed my view, the walls of my glass tunnel disappeared. I now live in the open air. There is still a difference between my life and the lives of other people. But the difference is less. Other people are closer. I am less concerned about the rest of my own life, and more concerned about the lives of others.

(1984, 281)

Having deflated the prudential and moral significance of identity, Parfit realised that our life will extend in an important sense through the lives of our continuers. What he might not have realised, though, is that if we'll continue to 'live in the open air' after death, we'll also remain susceptible to harms and benefits. It turns out that in this thought there is an account of posthumous harm which Parfit hadn't yet spotted, and I'm going to bring it out.

On my Parfitian view, it's not the deceased's *own* wellbeing level that is affected by a posthumous harm, because identity is not what matters to harms and benefits (in a sense that will be specified). The subject of harm is always a continuer of the deceased—a conscious individual who exists at the time of the harm and who is the bearer of wellbeing. It is a posthumous harm in the strict sense that an individual suitably psychologically related to the dead person is harmed. Call it a posthumous *quasi-harm* if you like, but for practical purposes it is relevantly like being harmed in a case of ordinary survival across time. Or so I'll argue (rejecting both *Identity-Tracking Harm* and the *Existence Requirement*).

This is a solution to the persistent puzzle of whether and how posthumous harms are possible that is consistent with all wellbeing theories, as it is silent on the nature of welfare goods. Surprisingly, it is even compatible with *experientialism*, the view that all quantitative differences in wellbeing

levels supervene on qualitative differences in conscious experiences, often associated with hedonism. I consider this a major advantage of the Parfitian view, as the traditional argument for posthumous harm is dialectically effective only for those who already align with the pre-theoretical judgments typical of desire-fulfilment theorists—a potentially narrow group of philosophers.¹³⁰ However, since most of what I'll say is intended to be compatible with all theories of wellbeing, I won't rule out that desire-frustration can in principle suffice for posthumous harm. I'll focus on sufficient conditions for posthumous harm rather than necessary conditions.¹³¹

I'll begin by considering some fission-cases to illustrate the Parfitian view. They'll give us reason to think that the holding of certain psychological continuities can make *us* prudentially concerned for the wellbeing of someone else. I'll then argue that such interpersonal psychological continuities can also hold in ordinary life and are indeed often preserved after death. I'll maintain that this implication of the Parfitian view explains the intuitive appeal of standard cases of posthumous harm. Finally, I'll defend my theory of posthumous harm against a number of objections and give some reasons to favour it over competing theories.

2. Fission-Cases and Posthumous Harms

You know that many bad things will happen after your death—to distant strangers, to your neighbours, to the people you care about most. But can any harm still happen to *you*? This question merits special consideration because

¹³⁰ Gardner, for example, resists the traditional desire-based theory of posthumous harm because she believes, as do many others, that 'an event cannot harm you unless there is at least one point in your life at which you experience its consequences for your health, your job, your reputation, your relationship with your family, or your relationships with your friends' (2022, 889).

¹³¹ In the same vein, desire-fulfilment theorists often commit themselves only to sufficient conditions for harm. Boonin, for example, understands the Desire-Fulfilment View as follows: 'If A's act frustrates B's desire, then A's act harms B' (2019, 11). The same spirit runs throughout this chapter, though I'll also examine possible necessary conditions for posthumous harm in Section 5.

your own wellbeing typically matters to you in a rather unique way: you can look forward to your own enjoyments and dread your own sufferings in a distinctively egoistic manner. And more abstractly, you might believe that *only* the wellbeing of individuals identical to you—*your own wellbeing*—deserves this special prudential concern. Let's call this the *Identity-Tracking View*.

In this section I'll argue that we should reject the Identity-Tracking View because of certain special cases. Consider:

Harming Your Fission-Offshoots. Your worst enemy plants a bomb in your car, causing an explosion that destroys your body. Miraculously, your brain remains intact, and the hospital has two corpses with functional bodies but destroyed brains. At T_1 , you undergo *Fission*: your brain's hemispheres are split and individually transplanted into these two new bodies. Each resulting individual—a fission-offshoot of yours—carries on all your pre-fission memories, intentions, beliefs, and the very same qualities of conscious experience, just as in ordinary survival across time. At T_2 , however, your worst enemy returns and tortures your fission-offshoots for the rest of their lives.

Have *you* survived *Fission* at T_1 ? Are *you* harmed by the torture inflicted on your fission-offshoots at T_2 ?

Insofar as there is one individual before and two after, and a determinate answer as to whether you survive *Fission*, we are left with three possible options: (1) you survive as one of the two resultant persons, (2) you survive as both, or (3) you don't survive. Assuming no relevant differences between your two hemispheres, it would be arbitrary to say that you survive as only one of the two individuals.¹³² It is also implausible to conclude that you

¹³² I am here following the classic Parfitian treatment of *Fission*, assuming, among other things, the falsity of a Lewis-style multiple occupancy view, according to which there are multiple overlapping lives at the pre-fission stage (Lewis 1976). One reason to be suspicious of this view is that it seems to entail an implausible overpopulation of welfare subjects, with potentially troubling implications. For example, if I am the kind of being who cannot undergo *Fission* while

are identical to both your fission-offshoots. For them to be identical to you, they would need to be identical to each other. But in principle, at least, they could occupy distant spatiotemporal locations, marry different people, or learn different languages. They are different persons, and to claim that you are identical to both would violate the transitivity of identity. Like Parfit, I accept the only remaining alternative as the best description available: you don't survive *Fission*.¹³³

Without survival at T_1 , what to make of the torture at T_2 ? According to the Identity-Tracking View, you shouldn't worry about it in any distinctively egoistic way. After all, those who are tortured are your fission-offshoots, and you are not one of them. But this implication seems implausible. If only *one* hemisphere of your brain had been transplanted into a body, you would have faced years of torture (people can and do continue to live after losing one hemisphere). Yet, even knowing this risk, it wouldn't have seemed rational for you to beg the surgeon to transplant *both* of your hemispheres. Breaking the logic of identity would technically save you from being the victim of torture. Even so, it wouldn't seem to make things any easier.

Identity-tracking theorists could try to defend their view by appealing to *reductionism*, according to which it is indeterminate whether you survive *Fission*, and so is whether you are harmed when your fission-offshoots are tortured, because there is no 'further fact' over and above the holding of certain physical or psychological continuities.¹³⁴ The original idea that 'you don't survive *Fission*' is just the 'best description' rather than the truth.¹³⁵

This won't help the Identity-Tracking View. On the above reductionist story, if you hadn't undergone *Fission* there would have been some future suffering that would have been determinately yours. But since you did

at some point you will, and if some aggregative moral principle is true, then it seems that torturing me would matter less, morally, than torturing you (or, more accurately, the multiple people who currently co-occupy your spatio-temporal location).

¹³³ Parfit (1984, 260).

¹³⁴ Parfit (1984, 259–260).

¹³⁵ Parfit (1993, 28).

undergo *Fission*, it is now indeterminate whether the suffering will be yours and whether you should worry about it at all. The problem with this story is that *Fission* is not intuitively a plausible example of ambiguous or indefinite prudential concern. You wouldn't be puzzled as to whether you should beg the surgeon to successfully transplant *both* of your hemispheres. It seems clear that you should regard the suffering of your fission-offshoots as if it were your own.

I believe we should reject the Identity-Tracking View and accept that the relation that underpins our prudential concern is something other than identity. But what exactly? Here comes the Parfitian proposal: this relation—*what matters in survival*—is the continuation of our mental life. Prudentially speaking, you should think of your fission-offshoots as if they were going to be you, because your mental states are psychologically connected with theirs via the same kind of continuities in experiences, memories, desires, beliefs, and character traits you ordinarily have with your future self. I'll call a *continuer* any individual who preserves this intimate psychological relation without identity. (I'll specify later the necessary and sufficient conditions for the holding of psychological continuities. For now, let's assume they hold at least in fission-cases.)

So far, so good: even if you don't survive *Fission*, or it's indeterminate whether you do, that doesn't matter. This is clear enough. What is not clear, however, is how the Parfitian view should understand the effect of harms to continuers on the wellbeing of the formerly living person: when a continuer suffers a harm, is it a case of posthumous harm, such that the suffering of the continuer at the time it occurs is bad for the formerly living person even though she no longer exists? Or is it a case of harm to another individual whom the formerly living person once had a prudential reason to care about, but that now doesn't affect her wellbeing in any way?

Parfit himself gave us no answer, but I think I have one compatible with the spirit of his view. We should answer the question of whether we are harmed when our continuers are harmed in the same way we should answer the

question of whether we have survived *Fission*. We are not literally harmed. We don't survive. But that doesn't matter. *Fission* shows not only that identity is not what matters in survival, but also that identity is not what matters to harms and benefits. Let me unpack this.

In my view, *Fission* teaches us that your own wellbeing isn't the only wellbeing you should be egoistically (or quasi-egoistically) concerned for. I call your *extended wellbeing* the class of wellbeing that matters to you in a special prudential way. It is helpful to distinguish between your personal wellbeing and your extended wellbeing because wellbeing, or welfare, is typically defined in terms of what is intrinsically good for *you*, or how well *your* life goes—definitions which presuppose identity. By contrast, your extended wellbeing is an expanded conception of the traditional notion; it can include the wellbeing of others as well as your own. It concerns what is intrinsically good for the collection of individuals you are prudentially concerned about. Which individuals? Parfit gives us all the elements for a *Continuity-Tracking View*: what makes the wellbeing of certain individuals your extended wellbeing is the holding of certain kinds of psychological continuities between you and those individuals.

Since harms typically involve identity, let me also distinguish between *harms* and *quasi-harms*. While a harm to you lowers the level of your own wellbeing—either at a given time or over your lifetime—a quasi-harm to you lowers the level of your extended wellbeing—either at a given time or over the relevant lifetime. (If we want to define extended wellbeing in terms of aggregate lifetime wellbeing scores rather than momentary wellbeing scores, the relevant lifetime will encompass both the life of the original person existing at T_1 and the lives of any continuers existing at T_2 . If something bad happens to a continuer at T_2 , it might not reduce the lifetime wellbeing score of the individual existing at T_1 , but it will reduce the score of her lifetime extended wellbeing, which encompasses both T_1 and T_2 .)

This is not a case of posthumous harm in the traditional sense. Since I am a different individual from my continuers, we cannot have the same life, and since my wellbeing concerns how well my life goes, we cannot have the same wellbeing. But I can and should think of the wellbeing of my continuer as an extension of my own wellbeing without identity. And I can and should think of harms to them as quasi-harms to myself. For everything that matters in the intrapersonal relation I have with myself in ordinary survival is contained in the interpersonal relation I have with my continuers.

We should be careful not to double-count such harms. Consider the standard intrapersonal case: typically, whenever I am harmed as an adult, my former self is quasi-harmed because my former self was prudentially concerned for the wellbeing of my later self. This doesn't make the harm twice as great. Rather, it's more like evaluating the *very same* harm from the prudential point of view of the same individual at different times—or, in fission-cases, from the prudential point of view of different individuals. The harm to the continuer is also a quasi-harm to the individual who is continued. The quasi-harm, then, is *time-relative*: it's a quasi-harm to a particular individual at a given time, and it's grounded in the particular prudential interests that that individual has at that time. And this matters: as we'll see in Section 4, the possibility of making things better or worse from the point of view of a particular individual at a given time can give us some special reasons for and against actions related to what we owe to that individual at that time.

Who is harmed, then, at the time my fission-offshoots are harmed? There is clearly a subject of harm: who suffers the harm is my continuer, a conscious individual who exists at the time of the harm.¹³⁶ But the same harm can be redescribed under different guises: it can be said that as the formerly living person I am quasi-harmed at the time of the harm, in that the level of my extended wellbeing, which survives in my continuer, is decreased.

¹³⁶ This shouldn't surprise us, as there are good reasons to believe that only conscious beings can be bearers of welfare goods or bads. See Chapter 1.

It might seem strange that something can affect the extended wellbeing of the formerly living person if she no longer exists; that something of her nonetheless *survives in* her continuer. But relations, properties, and parts of ordinary objects can survive in other objects all the time. Why should it be any different for people? Consider a ship that is destroyed at T_1 , but some of whose original pieces are used to build a new ship at T_2 . It is clear that the original ship no longer exists. It is not identical to the new ship. But there is also a clear sense in which *something* of the original ship does survive in the new ship. If, for example, I destroy the new ship, it is accurate to say that I have destroyed both the new ship and some pieces of the original ship. I have two subjects of ‘harm,’ metaphorically speaking.

In just the same way, something of me survives in my continuers after I die. What is it, and why is it so dear to me? It is not clear that one and the same token mental state is preserved in fission-cases, or that those mental states that are preserved are parts of me. I suspect that what survives in my continuer is a certain *relation* across mental states: my mental life being *continued* in the right kind of way. And relations survive the death of their bearers all the time. Napoleon may have cared a lot about his ship, and after his death that ship didn’t cease to be Napoleon’s ship—the ship that Napoleon used to own and care about. Yet that ship was never a part of Napoleon; it was merely a thing that stood in a certain relation to him. It is this relation that survives.

Since the relation that survives in my continuer is the kind of psychological continuity that grounds my special prudential concern for the future, it not only makes sense to say that the suffering of that continuer will affect my extended wellbeing—even though I will no longer exist—but it also makes sense to anticipate that suffering from within. When it occurs, it’s somehow relevantly like it occurs to me then. And my vicarious anticipation isn’t a total mistake. It is true that once I no longer exist, I will no longer experience any suffering. But my anticipation today remains rational, because

it remains true that the object of my prudential concern—the thing I care about—will outlast me.

Posthumous quasi-harms of this kind do not involve any backward determination of welfare value: it would be a mistake if the formerly living pre-fission person were to claim that her extended wellbeing level is *right now* being lowered by the torture of her post-fission continuers. Her extended wellbeing will survive in her continuers just as parts, properties, and relations of an old ship can survive in a new ship, and it will be affected at the time of the torture. After all it essentially consists in the wellbeing of her continuers, precisely what's affected at the time of the torture. This is a noteworthy departure from desire-based posthumous harms and arguably an advantage of my Parfitian view, because backward determination of welfare value does not occur in standard intrapersonal cases of harm (undergoing torture at T_2 isn't bad for me before it occurs).

What I'm proposing is to revise our notion of posthumous harm in terms of quasi-harms to the deceased. This is revisionary in many ways. For one thing, I'm denying that the deceased herself needs to be harmed in the traditional sense in order to suffer something *relevantly like* a posthumous harm (*Identity-Tracking Harm*). In at least some cases of posthumous harm, it is not the deceased herself whose wellbeing level is lowered; only her continuers suffer such a harm, and so only the level of her extended wellbeing is lowered. But that's enough, because the difference between harms and quasi-harms isn't important for the one who is continued. For another thing, I'm denying that the deceased needs to exist at the time of the harm in order to suffer this special type of harm (*Existence Requirement*). It is sufficient that something of the deceased survives, namely certain psychological relations. All in all, I'm denying both *Identity-Tracking Harm* and the *Existence Requirement* insofar as these conditions are expressed exclusively in the language of identity.

From now on, then, whenever I say that you are *posthumously harmed* when your continuer suffers, I'll merely be using a shorthand for saying that:

(i) the level of your extended wellbeing is decreased; (ii) any decrease in the level of your extended wellbeing is a quasi-harm to you; (iii) there is no important difference between harms and quasi-harms for the one who is continued (all else being equal); and (iv) this quasi-harm is posthumous from the point of view of your own lifetime.

3. The Continuity View of Posthumous Harm

In what follows, I'll argue that if our prudential concern tracks psychological continuity in a Parfitian fashion, our extended wellbeing turns out to be affected by many ordinary posthumous events. *Fission* was just the tip of the iceberg. The psychological continuities we have with others in ordinary life differ from those we have with our fission-offshoots in both degree and quality, but they can still explain many of our attitudes towards the dead and vindicate our intuitions about standard cases of posthumous harm. I'll first outline the necessary and sufficient conditions for the holding of interpersonal psychological continuities in Section 3.1, then I'll work out the details of a theory of posthumous harm based on these continuities in Section 3.2 and go on to examine some of its special features in Section 3.3. There are many objections that I'll set aside for the time being, so don't be alarmed if some steps in the argument seem to ignore complications. I'll address what I take to be the key objections in Section 4.

3.1 Interpersonal Psychological Continuities

According to Parfit, the relation that underpins prudential concern is the combined function of *psychological connectedness* and *psychological continuity*. Connectedness consists in the holding of direct psychological links between mental states. Parfit favoured the view that for a mental state A at T_1 to be directly connected to a mental state B at T_2 , it is necessary and sufficient that there be (i) relevant psychological similarity between A and B, and (ii) a causal

connection of *any* kind that is responsible for this psychological similarity.¹³⁷ For example, a later desire is psychologically connected to an earlier one if the later desire is causally dependent on and relevantly similar to the earlier one. Continuity is defined in terms of connectedness: it is the holding of an overlapping chain of direct psychological connections over time. Mental state A can be continuous with mental state Z even though there is no significant degree of connectedness between A and Z as long as each intermediate psychological link (B, C, D, etc.) in the overlapping chain is connected without gaps.

There is no canonical account of the notion of psychological similarity in the philosophy of mind, and the characterisation is usually left implicit. In my view, a psychological connection requires *relevant* similarity, which in part consists of the kind of qualitative similarity that holds when a mental state at T₂ seems to be a continuation of another mental state at T₁, and in part consists of those shared properties that make it appropriate to describe two sets of mental states as ‘similar’ in ordinary language more generally. Psychological similarity, understood in this way, is a matter of degree. A great variety of mental states can naturally follow from an antecedent mental state, but some of them just preserve a *higher* degree of similarity than others; they are more *like* the antecedent mental state in terms of those shared properties that make

¹³⁷ While Parfit (1984, 284–87) makes it clear that any causal link across mental states would suffice, the similarity requirement is hidden in the arguments. In discussing the psychological criterion, Parfit implies that changes in character, which are presumably qualitative changes in one’s mental life, disrupt connectedness (ibid., 206–7). In his discussion of the intrapersonal discount rate, Parfit assumes that there is less connectedness over time within a life (ibid., 303). But why? If connectedness is only a matter of causal or counterfactual dependence, why assume that my distant future self is less dependent on my early self than my mid-life self is? Only the fact that ageing typically involves progressively greater qualitative psychological changes explains this assumption. Parfit also introduces the 19th-century Russian as an example of substantial loss of connectedness (ibid., 327), where what separates the earlier and later Russian are qualitatively different political beliefs and associated commitments, ideals, and desires. It’s hard to explain Parfit’s various arguments without concluding that he took psychological similarity of some sort to be required for connectedness.

it appropriate to describe two sets of mental states as ‘similar’ in ordinary language.

On the Parfitian view, establishing a psychological connection also requires a causal link of *any* kind. Whether this causal link is unreliable or indirect isn’t important.¹³⁸ Suppose, for example, I step into an *unreliable* teleportation machine that disassembles my body into atoms and scatters them randomly across the universe. If, purely by chance, my atoms are teleported in the correct order, producing a duplicate of me, this process would preserve all the psychological relations that matter to me. Even an *indirect* causal interaction, such as exchanging thoughts through an online conversation, can reproduce some of my mental contents in others in the right kind of way.

From now on, I’ll use the term ‘psychological continuity’ in a special way, to include both the holding of direct psychological connections and the holding of an overlapping chain of such connections. I’m doing this because the distinction between connectedness and continuity will not matter much for my purposes, and because Parfit believed both these relations to be important (1984, 301). I’ll also accept Parfit’s necessary and sufficient conditions for a psychological connection to stick to the spirit of his view. Not everyone will come away convinced. But strictly speaking, my argument will not depend on accepting such conditions; it can apply to other analyses of the notion of a psychological connection in a structurally similar way, as long as they agree with Parfit that identity doesn’t matter.

Consider, for example, the view that genuine psychological continuity needs to be grounded in the physical substrate of the brain.¹³⁹ If my brain were disassembled and its tissues transplanted into the brains of a hundred other people, I would cease to exist. Yet *something of me* would survive. There might

¹³⁸ Of course, there are also controversial cases of causation, such as the butterfly effect in chaos theory, where it may seem unclear whether what I do now can affect the lives (and the mental states) of distant strangers in the right sense of causation. For my purposes I’ll understand a psychological connection as requiring *any cause that is clearly a cause*—any cause that is clearly and widely recognised as a case of causation.

¹³⁹ McMahan (2002).

be people who retain my memory of hiking in the Alps with my parents, and others who share my tastes in music, books, and films. And if these people's lives go badly, I would suffer a significant quasi-harm, albeit of a weaker sort than those possible in fission-cases. (Note that this might have practical implications, particularly in the case of future medical technologies involving the transplant of minded brain organoids.)

It's also possible to make sense of posthumous harms on other views, such as one on which psychological similarity is by itself sufficient for prudential concern.¹⁴⁰ But I have good reasons to restrict my attention to the Parfitian view: it is one of the most popular, one that Parfit himself took very seriously, and one which has a plausible way of explaining how the object of my prudential concern can survive disruptions such as fission, sleeping, waking, and the constant deaths of brain cells. At the very least, its implications deserve careful consideration. And even though I won't defend the Parfitian view against its competitors given the scope of this piece, I will argue that it can provide a compelling explanation of what's going on in standard cases of posthumous harms and benefits, and to that extent the view itself gains even further plausibility.

The most important feature of the Parfitian view for our purposes is that it allows for interpersonal psychological continuities to be easily established across individuals. With the people we spend most time with—relatives, partners, friends, and colleagues—we engage in ongoing causal interactions via joint experiences, discussions, arguments, and exchanges, shaping each other's minds. We are so permeable to each other's mental lives that almost any interpersonal activity can foster common memories, beliefs, and interests. Think of the holiday memories you have with your family; the unique brand of humour you have developed with your dearest friends; the musical taste you have started sharing with your partner; how many of your beliefs and ways of thinking were passed down from teacher to student; and

¹⁴⁰ See Kolak and Martin (1987), Elliot (1991), and Campbell (2005).

how many of your values belong to a culture or tradition you have been brought up in.

Parfit himself, among others, recognised this implication of his view.¹⁴¹ I find it quite intuitive. On one hand, we are willing to make sacrifices for our close ones that we wouldn't make for strangers, which can be explained by the fact that our close relationships involve psychological continuity—the same sort of relation that makes it rational for us to regard our own wellbeing as something of special significance. On the other hand, we care about our future self in a privileged way, which can be explained by the fact that our close relationships do not involve the same strength of psychological continuity that holds within a life. This view seems especially intuitive if we think that at least part of what grounds our special concern for our close ones is the vast number of shared experiences, memories, desires, and values, and that our extensive contact with one another causally contributes to these bonds.

To be precise, I only regard the holding of psychological continuity as *sufficient* for special concern, not *necessary*. Special concern can also be justified on other grounds, and notable and fine-grained differences between the ways we care about ourselves and the ways we care about our loved ones should be explained in part by other factors. Consider a mother whose child was taken away from her in the delivery room and who, despite years of trying, was never able to have her child returned or even meet the child. Although the interpersonal psychological continuities between her and her child are extremely limited or non-existent, she may continue to care deeply about her child's future. The mother might justify her special concern by appealing to the parental obligations she has to the child, her unfulfilled desire for a life with them, or the personal meaning she attaches to her experience of pregnancy,

¹⁴¹ Parfit calls the continuers we find in everyday life our 'close ties' (2011, 136). Brink (1990, 1997a, 1997b, 2021) has made extensive efforts to explain our special concern for those close to us in terms of our prudential concern for our continuers. Similar points have been made by Whiting (1986) and Jeske (1993).

among other reasons. I don't wish to deny these claims. Special concern without continuity can be perfectly rational.

Let me clarify another preliminary worry. Normally, I may not only give extra weight to my own interests but also discount them. For example, it may be permissible for me to choose to sacrifice two of my arms to save a stranger's single leg. This special permissibility is arguably grounded in the special psychological continuities holding between my present and future self. But surely it is not similarly permissible for me to choose to sacrifice the arms of a loved one to save a stranger's leg.

Fortunately, the Parfitian view doesn't have this implication. In cases of maximal continuity, such as with my fission-offshoots, it is indeed permissible for me to arrange the sacrifice of two of their arms for a good reason if I choose to do so. However, in cases of weaker continuity, such as with my friends and children, this is impermissible precisely because I am largely discontinuous with them; they have their own set of interests, values, and desires. Once a certain threshold of discontinuity is reached, my continuers gain an autonomy that is sufficiently distinct from my own to make it impermissible for me to discount their interests (although this might not be clear-cut in borderline cases).

The Parfitian view can deliver plausible results only if interpersonal psychological continuity is taken to come in degrees in some sense. Remember that I'm using 'continuity' here to refer both to the holding of direct psychological connections, which do come in degrees, and to the holding of an overlapping chain of such connections, which Parfit understood as a transitive, all-or-nothing relationship (1984, 205). There may be ways to model an overlapping chain of psychological connections as a matter of degree.¹⁴² But, for our purposes, we can simply acknowledge that the *strength* of psychological continuity is a matter of degree, and that the degree of prudential concern that individual A should have for individual B depends on

¹⁴² McMahan (2002, 50–51). See also Chapter 4 (Section 4).

this scalar property. I understand the strength of continuity between A and B as a function of the number of overlapping chains of psychological connections, the number of direct psychological connections, and the degree of similarity between the linked mental states.

3.2 Posthumous Psychological Continuities

Some interpersonal psychological continuities will persist after we die. These continuities can take many forms—common memories, shared intentions, retained character traits, inherited values, and ideas improved over time. They can be causally linked by media, such as books, paintings, or movies. But they don't have to be. Our teachings can live on in those we have conversed with, our sense of humour in those we have laughed with, and our traditions in those we have lived with. Because the resultant continuers can suffer all kinds of harms, and because something of a dead person—the psychological relation that underlies her special prudential concern—can survive in her continuers, there is an extended sense in which a dead person can suffer posthumous harm. I call this the *Continuity View* of posthumous harm, a view directly derived from the Continuity-Tracking View of prudence.

On this view, we can distinguish two types of posthumous harm (though the distinction can be reformulated to apply also to posthumous benefits):

Non-comparative posthumous harms. The obtaining of intrinsic negative effects on the extended wellbeing level of a dead person, or her having a net negative extended wellbeing level.

Comparative posthumous harms. The obtaining of comparative negative effects on the extended wellbeing level of a dead person, or her being totally deprived of a net positive extended wellbeing level.

Non-comparative posthumous harms are decreases in the extended wellbeing level of the deceased brought about by intrinsically bad states, such as having continuers who suffer pain for a certain period or whose lives as a whole are not worth living. Deceased parents, for example, can be non-comparatively posthumously harmed if their children are tortured.

Comparative posthumous harms, by contrast, are shortfalls in the extended wellbeing level of the deceased relative to what it would have been had the harm not occurred. These include cases in which the deceased is prevented from having an extended wellbeing level that would otherwise have been net positive. If I care about the survival of a painting I made, which contains what I've found valuable and meaningful in my life, and it is destroyed by a rival just after my death, then I am deprived of all the continuities I would otherwise have had. Perhaps I would have had none, or my continuers would have had a very low wellbeing level, and my rival hasn't deprived me of much. Or perhaps I would have had many well-off continuers, and my rival has stolen from me my only chance to preserve what mattered most in my continued existence.

It may be hard to see how I can be quasi-harmed by *not* having continuers—I don't exist, nor do the continuers I would have had, so how can I be quasi-harmed? In my view, the quasi-harm of being deprived of well-off continuers is relevantly like the harm of shortening one's good life. If someone had killed Aristotle, they might think 'Had I not killed Aristotle, he would have had many good things in the future, and now he won't because of what I did—that's comparatively bad for him.' And similarly, if someone had destroyed Aristotle's last manuscript, they might think 'Had I not destroyed Aristotle's last manuscript, he would have had many quasi-good things in the future, and now he won't because of what I did—that's comparatively quasi-bad for him.' I can legitimately form judgments of this kind without assigning actual wellbeing properties (for example, momentary wellbeing levels) to the deceased Aristotle. It's only true as a counterfactual claim that there would

have been a wellbeing level: the continuers of Aristotle *would have* had a net positive wellbeing level if the deprivation had not occurred. Given this, the possibility of comparative posthumous harm seems no more problematic than the possibility of the comparative harm of death—a harm that most of us already take very seriously.

It's worth pausing to consider two apparent counterexamples. Suppose that there are plenty of people who appreciate Aristotle's writings and are thereby psychologically connected to him by ideas, values, and hopes, but who all have bad lives overall. One might think the Continuity View entails that having such continuers quasi-harms Aristotle. Intuitively, though, it seems more like a benefit or an honour of some kind.

Conversely, suppose there are many people who appreciate the writings of Baristotle—a less talented brother of Aristotle—but who share with him only his laziness, boredom, fears, and traumatic memories of war. Even if these continuers have overall good lives, it's hard to see how Baristotle could regard this as a quasi-benefit.

We can deliver the right judgments in these cases. We only need to accept that the extended wellbeing of the deceased is composed not primarily of the continuer's overall or lifetime wellbeing level, but of the momentary wellbeing level obtained by the continuer *at the time* they are psychologically related to the deceased. The duration of these interpersonal psychological relations can vary greatly depending on the particular continuer under consideration, and our intuitive judgments will vary accordingly.

Consider first those readers of Aristotle who reproduce his thoughts in their own minds only for a short period. The psychological connections quickly fade away as soon as the readers stop remembering or actively engaging with Aristotle's thoughts. We might call such readers *short-term* continuers. Now, how good is the life of a short-term continuer of Aristotle at the time the psychological connections hold? Quite good, I suspect. The hours we spend entertaining Aristotelian ideas, values, and beliefs seem good for us. This is a

profound and creative intellectual activity that expands our knowledge and is enjoyable in its own right. It's a quasi-benefit to Aristotle because the momentary wellbeing level is positive: prudentially, it's relevantly like Aristotle himself living once again some blissful moments of his intellectual life.

The lives of the short-term continuers could be bad at other times or even bad overall, but the only post-mortem wellbeing level that matters to the deceased is that which is obtained by his continuers at the time they are psychologically related to him. The wellbeing of the very same individuals at other times doesn't matter, as identity doesn't matter.

We should understand the case of Baristotle in a similar fashion. Being lazy, bored, afraid, or traumatised is typically intrinsically bad for us. When the short-term continuers of Baristotle reproduce these mental states, it's prudentially relevantly like Baristotle himself living once again some bad moments of his life. Perhaps the lives of the short-term continuers are blissful at other times and overall well worth living, but if Baristotle is not psychologically related to them at those times, he will only be quasi-harmed by their negative momentary wellbeing levels.

Of course, having continuers doesn't always work like that. Consider a scholar who has studied Aristotle carefully for years, and who has come to think and see things just as Aristotle did—an individual whose Aristotelian continuities have largely shaped his character. He is a *long-term* continuer. It's not implausible to say that Aristotle would be quasi-harmed if such a continuer were tortured. Perhaps the continuer hasn't thought much about Aristotle's work lately, yet deep-rooted psychological continuities still hold beneath the pain. The continuer has become more and more like Aristotle, to the point that making him suffer is prudentially relevantly like making Aristotle himself suffer. A similar story can be told about Baristotle's long-term continuers. Insofar as they have overall blissful lives—full of art, love, friends, knowledge, and joy—and the psychological relations they bear to him remain unbroken,

then for the purposes of prudence it's as if Aristotle himself had gained new welfare goods on top of his old welfare bads.

I've been arguing, in short, that the extended wellbeing of the deceased is composed of brief momentary wellbeing levels in the case of short-term continuers, and more lasting momentary wellbeing levels (or possibly even lifetime wellbeing levels) in the case of long-term continuers. What else determines a deceased's extended wellbeing level? Beyond duration, there are three dimensions of variation in the post-mortem psychological continuity of an individual: the strength of the continuity, its quality, and the number of continuers. These dimensions fully determine the extended wellbeing level of a dead person and the significance of any given posthumous harm. Let me explain strength and quality now; I will turn to the number of continuers in Section 4.

Strength can vary independently of duration: a long-term continuer could be very weakly psychologically continuous with Aristotle, as in the case of someone who read the *Nicomachean Ethics* once yet remembers Aristotle's theory of virtue throughout her life. The wellbeing level of a continuer fully contributes to Aristotle's extended wellbeing level only if the continuer is maximally psychologically continuous with him, as in fission-cases. In all other cases, the contribution made by the continuer's wellbeing level should be discounted proportionally to the weakening of psychological continuity.

Consider, for example, the following case from Parfit.¹⁴³ You meet a terminally ill stranger and develop a strong concern for them to be cured, but you never hear of them again. It is hard to believe that you could be harmed in any way should this stranger die. The Continuity View can explain this: the tenuous continuities generated by a single meeting do not justify any significant prudential care. Suppose, by contrast, that you have instead spent years caring for the stranger, engaging in countless philosophical conversations, all of which resulting in significant interpersonal continuities.

¹⁴³ Parfit (1984, 151).

Suppose you then part ways with this stranger, never finding out whether they will be cured. Now it seems more plausible that you would be benefited, in some way, were this person cured. The Continuity View explains this intuition too: the strong interpersonal continuities built over years justify a significant degree of prudential care.

Perhaps we should discard all continuities below a certain threshold as insignificant, as continuities that are not strong enough to warrant noteworthy prudential concern. This would explain, for example, why it seems irrational to be prudentially concerned for the wellbeing of someone who quickly skimmed one of my articles. The strength of any psychological continuity, if present at all, might be so low that it should be discounted to zero.

Yet I find it plausible to believe that even tenuous psychological continuities such as those between an author and a reader of a book, or an artist and a viewer of an artwork, can generate notable prudential concern if they involve *qualitatively* important mental states. Consider cases like Aristotle or Botticelli, where most of the mental life of the formerly living individual is irreversibly lost. It wouldn't be surprising if certain philosophical thoughts were among the most central aspects of Aristotle's inner life, or if Botticelli's aesthetic appreciation of a certain kind of beauty was among the dearest things to him, such that the preservation of these venerable traits in well-off continuers could still amount to a significant posthumous benefit. In these cases, it seems that the quality of the surviving mental characteristics partly compensates for the weakening of psychological continuity.

Parfit himself recognised that it would be a mistake to assess the prudential value of my relation to a given continuer based solely on the strength or degree of our interpersonal psychological relations (1984, 299). If between me and my continuer there is one-tenth of the strength of psychological continuity that persists in ordinary survival, it does not follow that having this continuer is one-tenth as good as ordinary survival. I need to

know not only how closely I am related to my continuer but also the quality of what is continued—whether the features that are continued are good, neutral, or bad for the one who has them. Just as surviving isn’t good or bad in itself, having a continuer isn’t good or bad in itself either.¹⁴⁴ Whether it is good to continue depends on the wellbeing level that is continued, which in turn partly depends on the quality of the continuities themselves. Remember Baristotle!

3.3 Six Special Features

Before proceeding to the objections, let me briefly outline six special features of the Continuity View. In my view, all of them are virtues for a theory of posthumous harm.

First, it gives us a natural explanation of why damage to someone’s posthumous reputation—one of the most traditional examples of posthumous harm—is bad for the deceased in many of the relevant cases. If, after my death, an enemy were to ruin my reputation by cleverly forging documents filled with lies and manipulations, they would establish weaker psychological connections rather than stronger ones, or psychological connections of bad quality rather than good, thereby corrupting the overall strength and quality of the continuities I will have posthumously.

Second, and symmetrically, it offers a rationale for the common practice of honouring the dead through memorial works. We seem to have *reasons of remembrance*: reasons to remember our loved ones after their death, to preserve their pictures, visit their graves, write letters ‘in memory of’ them, and repeat old rituals we used to do together—all while balancing these activities against the other prudential and moral demands of everyday life.¹⁴⁵ In my view, these efforts can be seen as some of our most personal attempts to

¹⁴⁴ I am here considering an individual who merely survives in the absence of any candidate welfare goods or bads beyond their own survival. Such a life itself strikes me as neutral (Lee 2023b).

¹⁴⁵ Dorsey (2022, 133).

benefit a loved one who has passed away. We might not consciously realise it in everyday life, but this is one way to ground our *pro tanto* reason to remember people: continuing the deceased's mental life in us can be good for them in an extended sense.

To be precise, by remembering a dead person, I'm not necessarily harming or benefiting them; rather, I'm putting them *in a position* to be quasi-harmed or quasi-benefited. Whether my act of remembrance quasi-harms or quasi-benefits the deceased depends on my wellbeing level, as well as the quality and the accuracy of what is remembered. Suppose I recall only negative memories of them—their sufferings, vices, and regrets. Engaging in this sort of remembrance would renew and consolidate former sources of harm to the deceased. If my intention is to quasi-benefit them, I should engage in *good* remembrance: the sort of remembrance that involves memories typically associated with a positive momentary wellbeing level. What makes a wellbeing level positive will be specified by the right theory of wellbeing: if I am in great pain, do I still have any reason to remember my dearest friend who has passed away? I might if a non-hedonistic theory of wellbeing is true, such as a theory on which knowledge, love, or friendship are intrinsically good for the one who has them.

Third, my Parfitian view justifies the widespread belief that we can honour dead people by advancing some of their projects, and it does so without presupposing a conception of wellbeing on which desire-fulfilment or achievement is a welfare good. Imagine Sophia, a philosopher who spent ten years developing a theory about what matters most, often discussing it with her students. Unfortunately, just before completing her magnum opus, Sophia fell ill and died. Unwilling to let her ideas go to waste, her students decided to take up the task and finish her theory. They carefully considered her notes, recalled the long conversations with her, and filled in the gaps in her theory as best they could. They published a book dedicated to her, which became a bestseller. As I see it, Sophia has been posthumously benefited by her

students, who ensured the continuation of her most cherished ideas in the minds of many well-off continuers.

Fourth, the Continuity View gives us an explanation of how it is that certain things beyond our own felt experience matter, not impersonally, but in relation to us. If my child is murdered, I may be quasi-harmed even if I never learn of it. My life—or better, the chain of lives with which I am psychologically continuous—would be going much worse than I think.

This doesn't mean that the Continuity View entails a *non-experientialist* conception of wellbeing, on which there can be harms and benefits that don't correspond to differences in felt experience. It entails only that *quasi-harms* and *quasi-benefits* need not correspond to differences in *the subject's own* felt experience. Indeed, the fifth special feature of this view is that it is consistent with the following version of experientialism:

Extended Experientialism. Any difference in the extended wellbeing level of a given subject corresponds to a difference in the felt experience of those individuals with whom that subject is psychologically continuous.

While I won't commit myself to this requirement, it's worth stressing that the Continuity View is, to my knowledge, the only theory of posthumous harm compatible with some form of experientialism about wellbeing.

Sixth and last, this view can justify reparations to dead victims of wrongdoing. Consider the popular idea that present generations should make up for past injustices such as slavery. It may be difficult to see how compensation can be owed to dead people, who no longer exist, or to descendants who were neither directly nor indirectly harmed by the injustice, such as those who have perfectly good lives and who would not have existed had the injustice not occurred. While various solutions have been proposed,¹⁴⁶

¹⁴⁶ An example is Ridge (2003), who contends that since the victims of past injustices strongly desired their descendants to flourish, we can posthumously benefit the victims and satisfy our

my Parfitian view offers an alternative. If a presently existing person continues the values, beliefs, and desires of those originally wronged—by carrying on their culture and traditions, for example—it is appropriate to compensate her to the extent that her wellbeing constitutes an extension of the wellbeing of the originally wronged people. Benefiting such an intergenerational chain of continuers could be a viable way to repair past injustices.

4. Objections and Further Problems

Let's now consider some thorny questions raised by the Continuity View as a theory of posthumous harm: Is it extensionally adequate? Does it commit us to an excessively demanding past-oriented morality? How far into the future can a dead person have continuers? Should we aggregate posthumous harms?

Roughly, we can arrange common-sense cases of posthumous harm into five categories:

CATEGORY	EXAMPLE
<i>Harms to One's Loved Ones</i>	Torturing a partner, child, or friend of a dead person.
<i>Harms to One's Reputation</i>	Spreading false claims about a dead person so that others remember them in a worse light than they should.
<i>Harms to One's Intellectual Possessions</i>	Destroying a dead person's artistic works, books, or family pictures.
<i>Harms to One's Material Possessions</i>	Allocating a dead person's money against the directives of their living will.
<i>Harms to One's Body</i>	Disfiguring a deceased body.

duties of reparation by directly benefiting their descendants. This line of argument, however, is conditional on the Desire-Fulfilment View, with its own problems.

An extensionally adequate theory of posthumous harm need not provide a rationale for all these ways of harming a deceased person. It might instead give a principled explanation for why some of these putative harms do not constitute genuine harms, or it might be compatible with independent features that account for the types of harm it cannot explain on its own. We've already seen how the Continuity View captures our intuitive judgments about the badness of harms to our loved ones, to our posthumous reputation, and to our intellectual possessions. Let's see how it covers the remaining cases.

One might want a theory that accounts for the badness of harms to the material possessions or the body of a deceased person—a theory that explains how wasting their savings or disfiguring their corpse can harm them. The Continuity View naturally covers some of the relevant cases. If after my death a lawyer manipulates my will so that my inheritance goes to a stranger instead of my children, this is likely to make my continuers worse off. Indeed, any event that affects the body or material possessions of a deceased person is *comparatively* harmful whenever it makes their extended wellbeing level lower than it would otherwise have been.

Other cases, however, resist this treatment. Consider an act of necrophilia: if no one learns of it, it's hard to see how it could have a negative effect on the value of the deceased's post-mortem psychological continuities. Or consider a theft of authorship: if a manuscript stolen from a dead person were published under someone else's name, yet still resulted in just as many posthumous continuers for the real author as it would have under their own name—with each continuer being just as happy and just as strongly connected to them—it's hard to see how this deed could quasi-harm the deceased author in any way.

Fortunately, we can accommodate common-sense judgments about these cases. It's sufficient to recognise that the post-mortem frustration of the desires of the formerly living individual can lower their wellbeing level. After all, I'm developing a theory of posthumous harm from the point of view of a

particular metaphysics of persons, not a particular theory of wellbeing. For the same reason, I must also remain open to the possibility that hedonism is true, and that neither the act of necrophilia nor the theft of authorship I described could genuinely harm or quasi-harm the deceased. But this is fine by me: since we can freely combine the Continuity View with popular theories of wellbeing to handle all common-sense ways in which we can harm a deceased person or give a principled explanation for why some of these putative harms are not actual harms, I take it to be an extensionally adequate theory of posthumous harm.

Let's turn to the different issue of how far into the future a dead person can be posthumously harmed on the Continuity View. Psychological continuity tends to weaken over time. The surviving continuities of the deceased will weaken as the people they used to know also pass away, and the overall number and closeness of the bonds linking the deceased to their continuers might gradually decline, until they drop to (or asymptotically approach) zero. Is this a problem for a theory of posthumous harm?

It could be an advantage. Following this downward trend, posthumous harms would tend to become less frequent and severe as time passes. This would in turn explain why the normative significance of the deceased tends to decrease gradually over time—what has been called the *Fading Thesis*.¹⁴⁷ Standard formulations of desire-based posthumous harms do not accommodate this tendency. I take this as a point against them, since the Fading Thesis seems plausible in the relevant cases. It is hard to believe, for example, that if a Palaeolithic man had a strong desire for all snakes to die and vanish from the world, the wellbeing level he had back then would still hinge on whether snakes are around today.¹⁴⁸ Yet once we consider a desire whose fulfilment or frustration isn't inextricably tied to a particular time, and we accept that its fulfilment or frustration can affect an individual's wellbeing

¹⁴⁷ Dorsey (2022, 132).

¹⁴⁸ See van der Deijl (2021) for a similar argument.

level even after their death, it becomes quite unclear why the remoteness in time of that desire-fulfilment or desire-frustration should matter. The Continuity View can avoid this problem and other problems if we recognise that the history of interaction and influence responsible for psychological continuity produces stronger reasons when that history is recent, and weaker reasons as time goes on—not because the passage of time is intrinsically adverse to the persistence of continuity, but because the chains of psychological connections tend to decay as the history of the world unfolds.

This doesn't mean that posthumous harms are impossible long after our death. Suppose that someone demolishes the Sagrada Família 300 years after Gaudí's death. This would still be bad news for the extended wellbeing of the architect: he would lose what had been an extraordinary vehicle for his mental life, one just as effective 300 years after his death as it was during his lifetime. This is the case for famous authors, artists, scientists—and philosophers!—whose creations sustain enduring psychological continuities with posterity. But ordinary people too can extend their prudential interests well beyond their own lifetime by engaging in multigenerational cultural movements or family traditions that will transmit some of their core values and beliefs far into the future.¹⁴⁹

A different set of problems arises from the sheer number of dead people. There are roughly 8 billion people alive today, but by some estimates a further 100 billion have lived in the past. If we can quasi-harm dead people, and the quasi-harm resulting from our act generates a moral reason not to perform the act, then centuries of moral reasons seem to fall upon our shoulders: reasons not to harm the continuers of Aristotle, reasons not to move a painting from a museum to a private collection, reasons not to destroy the cherished photographs of our grandparents, and countless other reasons against acting. If we can also quasi-benefit dead people, and the quasi-benefit resulting from our act generates a moral reason to perform the act, then we

¹⁴⁹ See Chapter 4 for an extended version of this argument.

might also have reasons to read Aristotle, reasons to visit art museums, reasons to dwell on the most cherished memories of our grandparents, and countless other reasons for acting. We might then wonder whether the Continuity View demands too much of us.

This partly depends on our broader moral theory. If our aim is to maximise the total wellbeing of all sentient beings impartially, our obligations could be very demanding indeed: our own interests could matter very little in the grand scheme of things. But precisely because we would then be concerned only with the total wellbeing of everyone, impartially considered, we would not concern ourselves with the particular relations holding between a particular wellbeing and a particular individual at a given time—such as whether wellbeing X is of special prudential importance to individual Y at time Z—unless we had instrumental reasons to do so.

Let's illustrate this with an example. Suppose that my fission-offshoots and I undergo five iterations of *Fission*, each time splitting my brain and transplanting its pieces into new bodies. Each iteration applies to all the branches of this family tree. Assume also that each piece of my original brain can still carry my full mental life into the future. If one of the thirty-two fission-offshoots resulting from the final iteration is happy, there will be many people—including me—who split along the way to bring him about, who had a special interest in his wellbeing, and who will be quasi-benefited by whatever makes him happy. But it would be a mistake to think that the happiness of this one individual is as good as the aggregate happiness of many. The extended wellbeing of the formerly living individuals doesn't contribute to the total sum of wellbeing; rather, it's grounded in the particular prudential interests that each of them had in the wellbeing of their continuers. That is why, in this case, the very same wellbeing of the very same continuer coincides with the extended wellbeing of multiple people: just as many different people can want

the same candy, many different people can have a prudential reason to care about the same wellbeing.¹⁵⁰

So taking quasi-benefits and quasi-harms into account won't make your wellbeing-maximising moral theory any more demanding than it already is. If, however, you value the interests of particular individuals at a given time over and above the impartial maximisation of wellbeing, then taking quasi-benefits and quasi-harms into account will ground additional moral reasons for and against actions. If I owe money to someone before she undergoes *Fission*, I should return it to her fission-offshoots. If I have a moral obligation to benefit a particular individual from the past, I may be morally required to benefit her continuers. And if I have a moral reason not to inflict severe harm on an individual even when doing so would promote the greatest aggregate wellbeing, I might have to refrain from destroying the life's work of a deceased person even when such an act could greatly benefit others. The relation of a formerly living individual to each of her continuers carries most of the moral reasons that used to be grounded in identity, since it preserves much of what used to matter about identity.

Giving some weight to the extended interests of dead people won't commit us to an excessively past-oriented morality. Even if a quasi-harm generates a *pro tanto* moral reason, that reason can be outweighed by other reasons. And since the post-mortem psychological continuities linking dead people to their continuers are typically much weaker than those linking

¹⁵⁰ Beyond the notion of prudential interest and extended wellbeing, there are other moral concepts that might work in such a way that, even if the harm itself doesn't grow any greater, the number of people with something at stake does. Consider moral responsibility. Suppose that one million votes are necessary for the implementation of policy X, which will cause suffering to innocent people. Two million votes are cast and exactly 1 million votes are for X, so X is implemented. The vote of each one of the million people was necessary for the adoption of the policy: had any one of those people voted against it, X would not have been implemented. This seems to be a case in which each of the million is fully responsible for the harms caused by X. Voter V bears great responsibility for that harm, and the fact that 999,999 other people bear equal responsibility does nothing to diminish V's responsibility—for without V's vote, X would not have occurred. The harm caused by X is fixed, but the responsibility is multiplied a millionfold, at least on some plausible views of moral responsibility.

presently existing people to their future selves or to their contemporaries, the reasons generated by quasi-harms to dead people will typically be much weaker than those generated by comparable harms to presently existing people.

That said, absent sufficiently strong countervailing considerations, the pro tanto reason generated by a posthumous harm can make it the case that one is morally required to perform or to refrain from certain acts. And I suspect that, at least sometimes, the reasons against causing quasi-harms to the dead can outweigh the reasons against causing harms to the living. It's not intuitively implausible to suppose that Aristotle had a stronger prudential reason to avoid the posthumous destruction of most of his books than to avoid a severe physical injury during his lifetime, given the exceptional number, duration, and quality of the continuities he could have had.

Exactly how to aggregate the reasons generated by posthumous benefits or harms is hard to say. One major variable is the number of continuers. Is it better for you to have a billion happy readers who barely engage with your work, or a single happy duplicate who is maximally psychologically continuous with you? The answer depends, at least in part, on the right theory of aggregation—a notoriously difficult problem, and one that certainly cannot be solved here. Perhaps having many feeble continuers could never be as good as having a duplicate, or as ordinary survival, however great their number. Or perhaps there is after all some threshold number of continuers that would suffice, however tenuous their continuities.¹⁵¹ Instead of considering the merits of each view here—a discussion that would take us too far afield—let me simply invite you to take my Continuity View and modify it as you wish, according to your favourite theory of aggregation.

¹⁵¹ Gustafsson and Kosonen (2022) make an interesting case for a total prudential aggregation of this sort. While they restrict their discussion to fission-cases, their arguments could be suitably revised to apply to the continuers I've considered in this chapter.

5. A Comparison with Other Theories

So far, I have proposed and defended an alternative view of posthumous harm. I will now argue that my view is not merely an alternative to existing theories: it is superior to them, because it offers a clearer account of the metaphysics of posthumous harm and delivers better verdicts across the relevant cases.

Consider first the *Surviving-Properties View*, on which some properties of the deceased can continue to exist after death. Feinberg (1974, 1977), for example, holds that some of our interests survive our death in more than a metaphorical sense. The posthumous interests of the deceased, on his view, can be frustrated or promoted by subsequent events, and such interests are the ‘real subjects of harm.’ Other advocates include Sperling (2008), who takes the entity harmed in posthumous cases to be a human subject that does not exist ‘in a physical or material way,’ and Belliotti (2011), who posits the existence of a post-mortem entity through which actual elements of the ante-mortem person persist.

The problem common to all these views seems to be their metaphysical obscurity. What kind of thing are these surviving properties supposed to be? If they lack the capacity for consciousness and are not bearers of welfare goods or bads, how can they be appropriate subjects of harm? A Parfitian metaphysics of persons would seem to provide a more robust foundation for the Surviving-Properties View. There is nothing particularly strange about the existence of certain psychological similarities and causal connections between different minds. And there is a good reason why the individuals bearing such properties can be subjects of harms: they are concretely conscious individuals. Harm to them is also a quasi-harm to the deceased, but this is not too strange either. It only means that certain prudentially significant psychological relations of the deceased are continued and extended in certain suitably connected individuals.

Feinberg himself abandoned his own version of the Surviving-Properties View. He came to reject the idea that interests are themselves the true subjects of harm, since that suggests a ‘bizarre ontological reification, as if each interest were a little person in its own right’ (1984, 83). He turned instead to the Desire-Fulfilment View of posthumous harm.

One might complain, however, that this theory of posthumous harm relies on an unsatisfactory theory of wellbeing. Aristotle sounds right when he says that ‘desire is consequent on opinion rather than opinion on desire.’¹⁵² We desire to experience love and to create lasting works of art because we believe these things to be good; we do not believe them to be good because they fulfil our antecedent desire for them. To treat desire-fulfilment as the fundamental good-making property seems to reverse the proper order of normative explanation. The Continuity View is intended to offer an attractive alternative to those who find the possibility of posthumous harm plausible but remain unconvinced by a desire-based theory of wellbeing.

In what follows, I won’t discuss the problems that the Desire-Fulfilment View already faces as a theory of posthumous harm.¹⁵³ Instead, I’ll examine three overlooked problems that arise when we compare the Desire-Fulfilment View with the Continuity View. Our prudential concern and our desires often align, giving similar results—we normally desire things to go well for those entities for which we are rationally prudentially concerned. But there are limit cases where these two come apart. I’ll consider three types of cases: equally strong desires plus psychological continuity (5.1), psychological continuity without desires (5.2), and desires without psychological continuity (5.3).

¹⁵² *Metaphysics* (1072a, tr. Ross).

¹⁵³ Boonin (2019) offers the most extensive examination of these issues.

5.1 Psychological Continuity Plus Desire-Fulfilment

Consider two cases:

Grandmother's House. For her entire life, a grandmother took care of her childhood home. She designed the chairs, the living room, and the fireplace, and created all the paintings and decorations. She had a strong desire that her house outlast her—not because it was a masterpiece of impartial value (she was neither a great designer nor a great painter), but because she wanted others to live where she had lived her happiest days. She hoped they would continue sharing the background of her experiences, form somewhat similar memories, appreciate the products of her aesthetic imagination and sensibility, and partly feel what she felt during her life. After her death, however, a construction company demolished the house for no good reason.

The Villa. During his lifetime, a billionaire spent much of his fortune building a luxurious villa with a spectacular swimming pool. He rarely holidayed there and had very few personal memories associated with it; all his intellectual possessions, such as his books and pictures, were elsewhere. Yet he had a strong desire that the villa persist far beyond his lifetime, as a symbol of his wealth. After his death, however, a construction company demolished the villa for no good reason.

Let us assume that the grandmother and the billionaire desired the preservation of their possessions with equal strength, and that, all else being equal, they had equal levels of welfare. When their possessions are destroyed for no good reason, are they posthumously harmed to roughly the same extent, or is there a relevant difference?

Intuitions may differ, but despite the equal strength of desire-frustration, the grandmother seems to me to have had much more at stake than the billionaire. While the villa's walls and swimming pool would have failed to preserve the billionaire's values, beliefs, and memories, the grandmother's house would have allowed for importantly rich continuities. Her loss is then comparatively worse for her. I am not simply making the point that noble things are more valuable than material things for those who have them, but that desire-fulfilment plus continuity would have seemed much better for her than desire-fulfilment without continuity would have seemed for him. This demands an explanation. One plausible explanation is that the persistence of interpersonal psychological continuities with well-off continuers would have quasi-benefited the grandmother. The deprivation of this quasi-benefit is therefore an additional comparative loss for her over and above the frustration of her desire.

The desire-fulfilment theorist would get the wrong result if they understood desire-frustration as the only feature relevant to posthumous harms. They would have to conclude that the grandmother and the billionaire are posthumously harmed to roughly the same extent, thereby blurring the normative difference between the two cases. Of course, the desire-fulfilment theorist may object: who am I to claim that the continuation of our deepest values and memories matters more than the survival of a brand-new swimming pool? But if Parfit is right, it is simply a matter of prudential rationality that I should be concerned for the continuation of my mental life in a way that I shouldn't be concerned for the persistence of my material possessions, no matter how strongly I may desire otherwise.

If desire-fulfilment theorists want to accommodate our intuitions about these cases, there are a few moves available to them, but none entirely resolves the problem. The first is to revise their view and accept that only certain kinds of *informed* desires, formed under ideal conditions, are relevant to welfare. However, if we were to inexplicably form a preference for

posthumous psychological continuities over the preservation of mere material possessions after consulting an ideal advisor, then we would seem to be tacitly assuming a Parfitian account of prudential rationality. And from there, it is only a short step to accepting the possibility of posthumous quasi-harm and quasi-benefit.

A more promising move available to the desire-fulfilment theorist is to deny that desire-frustration is the only component of wellbeing relevant to posthumous harm. They could embrace a pluralist theory of wellbeing, on which not only desire-fulfilment but also knowledge, friendship, achievement, and other things are intrinsically good for one to have. Perhaps having continuers is one of these intrinsic goods, either because the continuation of our mental life is a distinct welfare good in and of itself, or because obtaining posthumous continuities counts as an achievement. This might do the trick. But the Continuity View still has simplicity on its side. It naturally follows from the internal resources of the Continuity View that having continuers can posthumously benefit us, while the Desire-Fulfilment View needs additional tools beyond desires to achieve comparable results.

The last move available to the desire theorist is to settle for a tie. They can acknowledge that their view, taken as a necessity claim, fails to fully capture the harm to the grandmother, but complain that the Continuity View too, taken as a necessity claim, fails to capture the harm to the billionaire. We should therefore settle for formulating both views as mere sufficiency claims; each is sufficient for posthumous harm in a way that does not exclude the other. This is the conciliatory spirit that runs throughout this chapter. In what follows, though, I briefly consider two cases that might challenge even this sufficiency claim. I treat these cases as exploratory, and not as undermining the conciliatory spirit of the chapter. It is up to the reader whether to give them any weight.

5.2 Psychological Continuity Without Desire-Fulfilment

We should consider cases of psychological continuity without desire-fulfilment to test whether continuity alone is sufficient for posthumous benefits or harms, even in the absence of conation.

Consider Van Gogh. He wasn't well known during his lifetime, yet his posthumous fame, especially among artists who deeply admired and studied his artwork, has made a major difference to the history of art. Millions of people now see Van Gogh's paintings daily, whether in museums, books, online images, or even on their smartphone and laptop backgrounds. Given how extensively Van Gogh's mental life is continued in the experiences of later admirers, particularly through perceptual and artistic experiences—aspects of his mental life that were especially dear to him—and given that those admirers would seem to have a positive momentary wellbeing level at the time they are engaging with his work, it seems that a very positive extended wellbeing has survived Van Gogh's lifetime, thereby quasi-benefiting him greatly. But let's assume, for the sake of argument, that Van Gogh never actually desired recognition for his life's work. Suppose he had no desire for his art to persist after his death or to be appreciated by others. He simply didn't care about having continuers. We would then have a case of posthumous benefit on the Continuity View but not on the Desire-Fulfilment View.

Speaking personally, I find it quasi-beneficial for Van Gogh to have people appreciating his work after his death, even if he may never have desired it. This intuition is strongest when I think of how many of Van Gogh's dearest imaginings and sensibilities may have survived in the imaginings and sensibilities of those who have dedicated their lives to investigating his artistic truths, how deeply these continuities may run across different minds, and how much these continuers may have benefited from the holding of such special connections.

Others may take this implication of the Parfitian view as a reductio. However, my primary concern here is with the conditional claim: *if* the Parfitian view is true, then we have grounds for posthumous harms and benefits of some kind. And this is harder to deny. Once we have accepted that Van Gogh's audiences really are his continuers, in the sense that their wellbeing is part of Van Gogh's extended wellbeing, Van Gogh would simply seem to be mistaken not to care about these individuals and his relationship with them.

Consider the following chain of cases. Most people would agree that even if Van Gogh didn't desire to have a happy life, he would nonetheless be benefited were he to have such a life. In the fission-case too, I suspect that most people would agree that even if Van Gogh didn't care about the welfare of his fission-offshoots, he would nonetheless be quasi-benefited were they to have happy lives. Now, if we take Parfit seriously, and if we therefore accept that Van Gogh's audiences really are his continuers, then we should similarly conclude that Van Gogh would be quasi-benefited were he to have happy continuers. Were Van Gogh to lack such a desire, he would simply be making a mistake.

To reject the Continuity View on the basis of a case like Van Gogh's would then be to reject the revisionary spirit of a Parfitian metaphysics of persons. That is a legitimate move: I have not offered an articulated defence of this metaphysics, and doing so would be beyond the scope of this article. Still, anyone who takes this line owes us a better account of the rational basis of prudential concern, which is no easy task.

5.3 Desire-Fulfilment Without Psychological Continuity

We should now consider cases of desire-fulfilment without continuity. This will test the intuitive plausibility of the view (stronger than the one I advocate) that continuity is necessary for posthumous benefits or harms.

Consider:

Post-Extinction Achievement. Russia and the United States are at war, while also competing to send the first autonomous rocket to the moon. Amidst the conflict, a succession of nuclear attacks by both countries annihilates all life on Earth; no one survives. By chance, a Russian rocket is launched just moments before the nuclear apocalypse. The scientist who built this rocket had a strong desire for it to land successfully on the moon. Although he is now dead, along with everyone he ever knew or cared about, the rocket makes it to the moon.

On the Continuity View, the extended wellbeing of the deceased can be affected only insofar as there are continuers. Assuming that no alien civilisation will ever appreciate the Russian feat, the deceased scientist cannot be quasi-harmed or quasi-benefited in any way *after* the nuclear apocalypse. The final quasi-harm has been the irreversible deprivation of all the continuers he would otherwise have had.

By contrast, on the Desire-Fulfilment View, the Russian scientist's life went more or less well depending on whether the rocket actually landed on the moon. After all, if some preferences can be fulfilled even after one no longer exists, there seems to be nothing to prevent a preference from being fulfilled in a world where *no one* exists. The Continuity View can accommodate this result if we understand it as a thesis about the *sufficient* grounds for posthumous harms and benefits, as I've done so far. But suppose we find it hard to believe that anything can be good or bad for anyone on a planet without conscious life—for example, because we accept a version of the experience requirement on welfare. Then the Desire-Fulfilment View gives the wrong verdict. Only the Continuity View can deliver the right result, provided it is understood as a claim about the *necessary* and *sufficient* grounds for posthumous harms and benefits.

6. Conclusion

Philosophers have long wondered whether we can still suffer harms after the irreversible loss of consciousness that accompanies death. Yet it has largely gone unnoticed that, if we deny in a Parfitian fashion that identity is what grounds special prudential concern, we can defend a complete theory of posthumous harm. This is what I have tried to accomplish.

I have argued that something of the formerly living person can survive in others—certain relations of her mental life, and, with them, her extended wellbeing. This special class of wellbeing can decline or prosper after death, for it is nothing over and above the wellbeing of all those conscious individuals who are psychologically continuous with the formerly living person. Such continuities can span centuries, like the oldest books; reach billions, like the greatest cultural movements; and travel as fast as information across the internet. Even more surprisingly, a theory of posthumous harm based on these continuities can accommodate many, if not all, of the standard ways in which dead people are thought to suffer harm and unlock new ways to benefit them.

This piece of writing itself can be understood as my own attempt to do something good for Parfit. Further developing some of his dearest thoughts and carrying on his mental life is my way of thanking him for what his work did to my life and to the history of moral philosophy.

Chapter 4:

Quasi-Surviving through Future Generations

This chapter argues that future generations can preserve what matters in our survival—the relation that grounds our ordinary prudential concern. My argument appeals to a special feature of Parfitian views of personal identity. If what matters is not identity but certain kinds of psychological connection, then intergenerational psychological continuities of memories, intentions, beliefs, and values can link together the lives of people existing at different times and provide grounds for expanding our egoistic preoccupations to include distant future people. I investigate the ways in which future generations can carry on our mental life after we die, and I conclude that we have good reasons to anticipate the persistence of these continuities. This means that our death may be less bad than we think, but also that the extinction of humanity could be worse.

1. Introduction

I have finished a monument more lasting than bronze, more lofty than the regal structure of the pyramids, one which neither corroding rain nor the ungovernable North Wind can ever destroy, nor the countless series of the years, nor the flight of time. I shall not wholly die, and a large part of me will elude the Goddess of Death.

Horace, Book III

Many believe that they may survive death, in some sense, if future generations will read their books, embrace their ideals, advance their projects, or continue their traditions. Is there any justification for this widespread belief, or is it an instance of wishful thinking? Will I live on through future generations? Some philosophers believe that I might.¹⁵⁴ Others deny that this kind of survival could be possible in any meaningful sense.¹⁵⁵

Although future generations might not give us immortality, I argue that they can be our *continuers*, in the sense that their existence can preserve at least part of *what matters in our survival*. That is, their existence can preserve at least part of the relation that grounds the special, prudential concern we ordinarily believe we should have only for ourselves. I conclude that if

¹⁵⁴ Johnston maintains that at least morally good people can genuinely survive through ‘the onward rush of humanity’ (2010, 331). Altshuler (2017, 2021) argues that certain psychological connections between our mental states and those of future people can extend our own existence well beyond our lifetime.

¹⁵⁵ Scheffler, for example, argues that we have a desire to preserve our social identity after we are gone in a way that doesn’t involve ‘the false belief that one will actually survive one’s death’ (2013, 34).

humanity does not go extinct, we can plausibly expect a significant degree of what matters to be passed on through the generations to come. My conclusion has explanatory and normative implications. It provides an account of why we are already distinctively concerned for future people, and gives us a reason to be even *more* concerned for them.

2. Quasi-Survival

Consider two stories:

Grandmother. A grandmother shares all her cherished experiences, beliefs, values, and aspirations with her grandchildren. By doing so, she says, ‘I’ll live on through their memories.’ More than that, she hopes that some fragments of herself will go on exerting an influence on posterity long after she is gone—from her grandchildren to her unborn great-great-grandchildren, and so on down the chain of generations.

Writer. A writer believes that every time a reader reads his books and appreciates them appropriately, they in part think what he thought, feel what he felt, and grasp what he grasped. ‘Through the act of reading the lines I wrote,’ he explains, ‘the patterns that form their thoughts come to imitate the patterns that once formed mine. The feelings that fill their minds begin to resemble the feelings that once filled mine. In those moments of literary appreciation, then, I live on through the experiences of my readers to the extent that their experiences mirror mine.’ He repeats to himself that, if he makes it as a writer, he’ll live on through the experiences of people far beyond his own lifetime.

What should we say about these stories? Much like *Grandmother* and *Writer*, many of us care about the posthumous traces we leave to posterity. But, taken

literally, the claims of *Grandmother* and *Writer* can be understood as something stronger: as aspirations for individual survival after death. And appealing as this route to immortality might be, there are strong grounds for scepticism. For one thing, it seems implausible that one's identity could split into a one-to-many relation, so that one and the same person could be numerically identical with multiple people existing simultaneously in the future. The transitivity of identity would be violated: for multiple future people to be identical to me, they must be identical to one another. But these people will all exist in different places and lead different lives. They will have different bodies, love different partners, speak different languages. And if they are not identical to one another, how could they all be identical to one and the same person, me?

Perhaps, however, when *Grandmother* and *Writer* say 'I will live on' they don't mean this literally. It may be a manner of speaking, a metaphor meant to capture something of importance. I want to take up this line of thought and suggest that we should frame the question 'Will *I* live on through future generations?' differently, removing any reference to identity. What we should be asking is: 'Will *what matters to me in my ordinary survival* live on through future generations?' While I do not believe we can achieve literal survival through our descendants, it does seem to me that at least part of what matters in our continued existence might persist through the stages and changes of the future history of humanity.

This distinction between personal identity and what matters in survival was first drawn and famously developed by Derek Parfit. Parfit's thought was that we can treat the questions 'Will that individual be *me*?' and 'Will that individual have what matters to me in my ordinary survival?' separately. To clarify the distinction, consider a machine that scans your body, destroys it, and then creates an exact replica of you, identical in all your physical and mental properties. Before undergoing the process, you may ask yourself two questions:

Q1. Will I be this future person?

Q2. Will this future person have what I care about in my ordinary survival?

Even if this future person may not literally be you—perhaps because your survival requires the persistence of your body—it may nonetheless be rational for you to be prudentially concerned for them in the same distinctively egoistic way you are ordinarily concerned for yourself. It may be rational for you to anticipate their experience and their welfare: to hope that they will have joyful days and to dread the prospect of their suffering. In that case, this person would have what matters in your survival: the relation that underlies the special, prudential concern we ordinarily believe we should have only for ourselves, and that makes our own survival matter to us the way it does. Even if you wouldn't survive this machine, then, you shouldn't necessarily despair at the prospect. You would still *quasi-survive* were what you care about in survival to persist without identity.

Though literal survival through future generations is unattainable, it remains an open question whether we may quasi-survive through them. Identity holds only one-to-one; by contrast, the what-matters relation could hold one-to-many, from me to multiple future people. Even if I die now without an exact duplicate, part of what matters in my continued existence may persist through those who share my memories and values, or those who carry on my intentions and projects. The thought can be liberating, consoling even. Parfit himself came to regard his own death as less bad in light of these interpersonal relations:

[After my death] there will later be some memories about my life. And there may later be thoughts that are influenced by mine, or things done as the result of my advice. My death will break the more direct relations between my present experiences and future

experiences, but it will not break various other relations ... Now that I have seen this, my death seems to me less bad.

(1984, 281)

For Parfit, death disrupts the closest connections with our future self, but other psychological relations, causally arising from our life, will continue through other people.¹⁵⁶ These relations matter. So, our death doesn't amount to the loss of all we care about in survival.

This fascinating idea has received relatively little critical discussion.¹⁵⁷ I want to develop Parfit's insight more thoroughly and see whether the kind of quasi-survival he describes can reasonably be expected to occur. I will argue that our relations to future generations can in fact be expected to preserve a significant degree of what matters.

My conclusion has revisionary implications. Many believe we have moral reasons to worry about future people. But if I am correct, then we would also have strong *prudential* reasons, and thus stronger reasons overall, to ensure the survival of humanity.¹⁵⁸ This intergenerational prudential concern may directly affect our rational decision-making and, in turn, impact the long-run trajectory of human civilisation. Consider the following example. You are prudentially required to invest your disposable income if you are virtually certain of large returns next year. Likewise, if prudence demands a long-term view (given a reasonable chance of quasi-survival), you may be prudentially

¹⁵⁶ Later in *On What Matters* (Vol. 1, 136), Parfit says that the interpersonal psychological relations we bear to our 'close ties,' such as our close friends and those we love, are to some degree analogous to the intrapersonal psychological relations that ground special care for ourselves.

¹⁵⁷ The only work that I am aware of is Altshuler's (2017, 2021), which offers an intriguing, albeit somewhat preliminary, defence of this Parfitian approach. My project aims to work out this view in proper detail.

¹⁵⁸ Gustafsson and Kosonen (2025) also argue that we have prudential reasons to care about the far future of humanity, but their case appeals to the astronomical expected value of anti-ageing and mind-uploading technologies. I'll argue that prudence demands a long-term view even without relying on such technological developments, just in virtue of the right account of what matters.

required to invest significant sums to promote humanity's long-term flourishing. Prudence might therefore incentivise us to devote more of our personal resources to benefit the far future.

We may also have other egoistic grounds for doing this. We may desire the distant future to go in a certain way, or have plans and projects whose realisation depends on how they are continued by future generations. And because the fulfilment of such desires and plans might be important to how well our own lives go, we could care about the future of humanity for our own sake. I won't be denying that these things might matter too. But I'll be arguing that we can have prudential reasons to care about the future of humanity just in virtue of what matters in survival. An advantage of this argument is precisely that it is compatible with any theory of welfare, even those on which the post-mortem fulfilment of desires, plans, or projects doesn't matter for welfare. In fact, even hedonists would have reason to care about the pleasures and pains of far future people, so long as those people will be psychologically related to them in the way I'll describe.

3. Two Strategies for Quasi-Survival

Instead of asking whether I shall live on through future generations, I ask whether my relation to them will contain what matters in my survival (henceforth simply 'what matters'). But what must future generations be like to preserve what matters?

The common-sense view is that my prudential concern is rationally justified for myself, and myself only. If this *Identity Account of What Matters* is true, future generations will not preserve anything that matters in my survival. They will not literally be me. Parfit, however, makes a strong case against this view. Suppose you undergo *Fission*: your cerebral hemispheres are individually transplanted into two separate bodies.¹⁵⁹ Each resulting individual

¹⁵⁹ Parfit (1984, 254–255).

carries on your pre-fission memories, intentions, beliefs, and values, just as in ordinary survival.¹⁶⁰ What has happened to you? You cannot be just one of the two resulting people. It would be arbitrary to suppose that one of them is you but not the other. You cannot be both people, as the transitivity of identity would then require these two people to be the same person—you. But they are distinct individuals. They may travel to different places, enjoy different streams of consciousness. The remaining alternative, Parfit argues, is that you do not survive *Fission*. Yet this is no ordinary death. You are not there, but nothing is lost. Your fission-offshoots preserve all that matters to you. In fact, even though you would remain yourself if you survived with only one hemisphere, it would be irrational to prevent *Fission* by, say, begging the surgeon to destroy one hemisphere before the transplant.

Parfit proposes a *Psychological Account of What Matters*, according to which what you should really care about is not that *you* will exist tomorrow, but that tomorrow there will be someone who continues your mental life. More precisely, what matters is the combined function of two relations: *psychological connectedness* and *psychological continuity*. Connectedness consists in the holding of direct psychological connections between mental states. A later desire, for example, is psychologically connected to an earlier one if the later desire is causally dependent on, and relevantly similar to, that earlier desire (the precise characterisation of similarity across psychological mental states is a debated issue in philosophy of mind and I'll leave this notion implicit; though see Chapter 3, Section 3.1). Continuity is defined in terms of connectedness: it is the holding of an overlapping chain of direct psychological connections over time. Mental state A can be continuous with mental state Z even where there is no significant degree of connectedness between A and Z, as long as each intermediate psychological link in the overlapping chain is

¹⁶⁰ There are many real-world cases of people who go on to lead a normal, or nearly normal, cognitive life after losing one hemisphere.

connected without gaps (A to B, B to C, C to D, and so on). Both relations can hold one-to-many across multiple individuals.

Despite the proliferation of views of what matters that have followed Parfit's proposal,¹⁶¹ the Psychological Account remains one of the most compelling and widely embraced. If Parfit is right, our relation to future generations will contain what matters to the extent they will continue our mental life. Admittedly, this hypothesis is still too general to be informative. We must ask what kind of causal connection is needed to link mental contents across time in a way that preserves the object of our prudential care.

There are several alternatives. On the *Narrow View*, what matters is preserved if, and only if, our mental states are causally connected in the same way they are ordinarily connected via our brains.¹⁶² But the way our mental states are connected intrapersonally, via our brains, is different from the ways in which mental states can be connected interpersonally across different people. So, if the *Narrow View* is true, the existence of future people cannot satisfy the what-matters relation. This view faces problems—in my view decisive problems—but I won't rehearse them here.¹⁶³ Instead, I'll focus on views on which what matters can be preserved through connections other than brain-based ones, and ask whether such connections can allow for what matters to persist through the future of humanity while delivering intuitively plausible results.

Take the *Widest View*, favoured by Parfit.¹⁶⁴ On this view, for a mental state A to be connected to a mental state B, it is sufficient that there be (i) psychological similarity between A and B and (ii) a causal connection of *any*

¹⁶¹ For example, Unger (1990), Brink (1997a), and McMahan (2002).

¹⁶² Parfit (1984, 207–209).

¹⁶³ Parfit considers the case of a person who uses computerised artificial eyes to obtain visual experiences. These eyes, he says, could in principle preserve what we care about in vision without involving any ordinary brain-based connections (1984, 207). He takes this case to show that when it comes to prudentially relevant connections between mental states, we do not care about the particular substrate or medium.

¹⁶⁴ (1984, 207).

kind responsible for that similarity. An unreliable or indirect cause can preserve the object of my prudential concern.¹⁶⁵ Suppose, for example, that I step into an unreliable teleportation machine, which destroys my body and scatters my atoms across the universe. Suppose further that my atoms are teleported and reassembled in the right order, producing a duplicate of me purely by chance. This duplicate would still have what matters to me. An indirect causal interaction between my mental states and those of others, such as exchanging thoughts through a conversation, can likewise reproduce some of my mental contents in another person. Crucially, I may reproduce my existing psychological traits in people who will live beyond me, who may then reproduce those same traits in their successors. Call this the *Continuity Strategy* to quasi-survive through future generations.

There is also another strategy for quasi-survival. This is based on a view of what matters that is even wider than the one Parfit called ‘the widest.’ On this view, no causal connection is required at all. We need only a similarity requirement, according to which what matters is the persistence of psychological and phenomenal characteristics qualitatively similar to our own.¹⁶⁶ This similarity relation comes in degrees, and it is reasonable to expect future people to be like us in many mental respects. Call this the *Similarity Strategy* to quasi-survive through future generations.

We thus have two strategies for quasi-survival. On the Continuity Strategy, what matters is the extent of intergenerational psychological continuity and connectedness maintained across the future of humanity, which depends in part on causal relations holding between mental states. On

¹⁶⁵ This contrasts with the *Wide View*, on which the right causes are reliable ones (Parfit 1984, 207). I set this view aside, since I am not aware of any proponents.

¹⁶⁶ McMahan (personal communication) reports that Parfit later in life came to believe that psychological similarity between mental states is sufficient to preserve what matters in survival even without any causal relation. Although I find this reading of Parfit attractive and worth further investigation, in this thesis I confine myself to the official statement of the Parfitian view as presented in *Reasons and Persons*. For other proponents of a similarity-based view of what matters, see Kolak and Martin (1987), Siderits (1988), Elliot (1991), and Campbell (2005).

the Similarity Strategy, what matters is the degree of psychological similarity between us and future people, regardless of how that similarity is obtained. For the remainder of this chapter I'll address only the Continuity Strategy, in order to work out the view in sufficient detail; the Similarity Strategy will be examined in future work.

4. The Continuity Strategy

Before considering why we should extend our prudential concern to people who do not yet exist, I must explain why we should extend it to others at all. The first step is to accept the Widest View. There are good reasons to do so: it offers a credible account of how the object of our prudential concern can survive disruptions of various sorts, such as the branching of fission, the interruption of consciousness in sleep, and the constant replacement of neural tissue. It is also a popular view, and one that Parfit himself took seriously. Not everyone will come away convinced, but at the very least its implications deserve close attention.

Once we have accepted that a psychological connection between mental states requires only psychological similarity and a causal relation of any kind, we can see how easily such connections are established across different individuals. As human beings, we are constantly engaging with others, shaping one another's mental lives. With those we are closest to—relatives, partners, friends, and colleagues—we have ongoing causal interactions through joint experiences, conversations, and collaborative work on common projects. We come to share interests, memories, tastes, and values, developing deep psychological similarities. For the purposes of prudential relations, we are connected to others in ways analogous to those in which we are connected to our own past or future selves. The difference between the extent to which we care about ourselves and the extent to which we care about others can be

explained by the differing strength of psychological connectedness and continuity, among other factors.¹⁶⁷

I find this implication plausible and intuitive. The special concern we reserve for the interests of our friends is relevantly similar to the special concern we reserve for our own interests. Many of us think, for example, that a benefit to a friend can compensate for a sacrifice of ours in a way that a benefit to an unrelated stranger cannot. If we admit that psychological connectedness and continuity ground special concern, we have a compelling explanation of this analogy.¹⁶⁸ This is especially plausible if we believe that at least part of what justifies concern for our friends is the fact that we share experiences, desires, and values with them, and that our interaction with one another is causally responsible for the formation of these bonds.

To some, this may look like a grim view of our special relationships. It may seem as if I am treating our reasons to love as egoistic reasons in disguise.¹⁶⁹ But that is only half the truth. On the Widest View, the distinction between self-concern and other-regarding concern collapses at a deeper level. If it is true that we care about our loved ones because they carry on our mental life, then it is also true that we care about ourselves because we carry on the mental lives of our loved ones. We can say that we care in a special way about others for our own sake just as much as we can say that we care in a special way about ourselves for the sake of others. Either description is as accurate, and as inaccurate, as the other. The object of our special care, in fact, is neither ‘us’

¹⁶⁷ See Chapter 3 (Section 3.1), where I consider a mother whose child has been taken away from her just after birth. Although the psychological connections between them are effectively nil, she may continue to care deeply about that child throughout her life. In fact, her concern may be grounded in the parental obligations she has towards the child, in the desire for a life she could not have, or in the personal value she attributes to the experience of pregnancy, among other things.

¹⁶⁸ Whiting (1986) and Brink (1997a, 1997b, 2021) are the leading proponents of this argument. Both seek to explain our special concern for those close to us in terms of the holding of interpersonal psychological connections.

¹⁶⁹ The ancient Greeks might have found this view congenial. Brink (1997a) argues, fairly convincingly, that the Platonic and Aristotelian doctrine of a friend as ‘another self’ anticipates the Parfitian idea that there can be psychological connections between people.

nor ‘them,’ but the whole chain of lives linked together by impersonal psychological connections.

The step from an interpersonal extension of the what-matters relation to the Continuity Strategy is a small one. The causal chain that connects mental states across different individuals does not stop at the people who currently exist but will continue its course far beyond them. After our personal extinction, our intimates will establish prudential relations with the following generation, and that generation with the one after, and so on indefinitely through the history of humanity. To the extent that what matters in our survival will be inherited by distant future generations, we should include them within the scope of our prudential care.

Let us call these future people, who would preserve the what-matters relation in the absence of identity, our *continuers*. There are two routes to continuers: one through psychological continuity, the other through connectedness. I will assume, in a Parfitian spirit, that each relation suffices on its own to preserve at least part of what matters.¹⁷⁰ (I see no strong reason to think otherwise, though depending on one’s intuitions about some of the cases I’ll give later, one might be inclined to assign greater weight to one relation over the other.)

I’ll start with continuity. Suppose G_1 , G_2 , and G_3 are consecutive generations of people. Connectedness consists in direct psychological relations between members of G_1 and G_2 , G_2 and G_3 , and so on. Adjacent generations can be connected by direct interactions such as those between parent and child, or teacher and student. There may be no direct psychological relations between members of G_1 and G_{99} , since at the time of G_1 the members of G_{99} did not yet exist. By contrast, psychological continuity—the holding of overlapping chains of such relations—can extend far beyond the circle of those with whom we have direct interactions. Continuity is a transitive relation: if members of G_1 are psychologically connected to members of G_2 , and those

¹⁷⁰ (1984, 301).

members of G_2 are connected to members of G_3 in the same mental respect, it follows that members of G_1 are psychologically continuous with members of G_3 .¹⁷¹ Generations that are not directly connected can nonetheless be continuous, provided that the members of each generation in the chain are connected to some degree with their neighbours without gaps (I'll occasionally speak of 'continuity across generations' for brevity, but connections always hold between individuals, not between groups).

There is also a way for members of G_1 to have some degree of connectedness with members of G_{99} . Even in the absence of in-person interaction or overlapping chains of connections, psychological similarity between mental states can be established through indirect causal links provided by some medium such as a book, a painting, or a sculpture. Suppose that G_1 is Classical Greece, and G_{99} is the Renaissance generation rediscovering the teachings and values of G_1 through the study of G_1 's artefacts. Suppose also that those values had been lost since G_{40} , the fall of Rome. If we apply the Widest View to this simplified example of history, it follows that a degree of intergenerational connectedness was renewed through the causal links provided by Classical Greece's artefacts. And this occurred whether or not continuity also held throughout the intervening centuries.

The legacy of what matters across generations differs from one individual case to another in how tight the bonds are and in how much direct interpersonal interaction they involve. It can go through stages of discontinuity and then reconnect again. Or it could break irreversibly at some point. It would presumably come to an end with the extinction of humanity, unless we were succeeded by a species or an artificial intelligence sufficiently psychologically similar to us to allow for bonds of connectedness. And it is far

¹⁷¹ Parfit defines continuity as a transitive relation in order for continuity to be a possible criterion of personal identity. Since personal identity is transitive, continuity must be as well (1984, 206–7).

from certain that humanity has a future.¹⁷² Still, it seems plausible that, as long as humanity survives, many of us will be significantly continuous with future people, given the transitivity of continuity. Or so I'll attempt to argue.

This conclusion, however, should not make us indifferent to our own death. Even for the luckiest of us, only *a degree* of what matters will persist through the cohorts of future people, not *all* that matters.

There are two reasons for this. Firstly, connectedness is one element of the what-matters relation. Picture the succession of generations as a family tree, with us at the roots and generations farther along the branches the more distant they are in the future. Since the strength of connectedness decreases the farther we move from the roots, we may be rationally justified in having greater concern for the nearest generations, with whom we have direct relations, than those generations that will exist in the more distant future. We obtain an intergenerational discount rate on prudential concern to the extent that connectedness has independent importance.¹⁷³

Secondly, my future continuers will not be continuous with me to the *same degree* that I am with my own future self. Continuity comes in degrees, as the number of overlapping chains of connections and their strength can vary (more on this shortly). The connections between each link in the intergenerational chain will hold to a lesser degree than those between me and my future self. For these reasons, having continuers among future generations should not make me indifferent between living and dying. My death is still bad, only less bad than I thought.

¹⁷² Ord (2020) estimates there is a one in six chance that humanity will experience an existential catastrophe within the next 100 years.

¹⁷³ This differs from the discounting of the future that Sidgwick (1907, 381), Ramsey (1928, 543–4), Rawls (1971, 293), and Parfit (1984, 125–6) argued against. It would be a discounting of the *prudential* relevance that future people's interests should have *for us*, not of their independent *moral* relevance. Of course, one could appeal to prudential reasons of this kind to justify the moral permissibility of some degree of partiality towards the interests of the nearest generations. But this step from prudence to morality might have troubling implications in the long run. It risks making it permissible, for instance, to impose vast harms on untold numbers of future people in order to spare a friend a small one (Mogensen 2022).

Another view might be that continuity is an all-or-nothing relation, perhaps because there is a certain number of psychological connections that must hold at every point in the overlapping chain in order for someone to qualify as a continuer—say, around 50% of the number that hold from day to day in the life of a normal person.¹⁷⁴ But any such threshold, whether precise or vague, strikes me as implausible. What if someone had 49%, 30%, or 10% of the number of psychological connections that normally tie a person together over time? Would this person not be a continuer at all?

Consider a case of late dementia, where each link in the overlapping chain of connections holds much more weakly than in a normal person, and where it's unclear just how weakly. We should not say that a person in the early stages of dementia has no prudential reason *at all* to fear the torture of his body in a later stage. So rather than taking continuity to require sufficient connectedness, I think we should take it as the transitive closure of connectedness-to-any-degree, in such a way that even weak overlapping chains of connections might preserve at least some of what matters. Continuity itself can then be understood as a matter of degree. The degree of psychological continuity is a function of both the number and the strength of the overlapping chains of connections, where the strength of each chain depends on the degree of connectedness at each point in it. I understand the Continuity Strategy accordingly, as committed to the thesis that we ought to be prudentially concerned for future generations proportionally to the degree of continuity and direct connectedness in which we stand to them.

Let me say a little more about how continuity can be both transitive and a matter of degree. Consider the following family of continuity relations. Three individuals—A, B, and C—are members of G_1 , G_3 , and G_6 respectively.

¹⁷⁴ This is Parfit's own view (1984, 205). He held that continuity is an all-or-nothing relation that requires sufficiently strong connectedness at each link. It is, however, unclear why continuity could not be a matter of degree even if it did require some minimum degree of connectedness in a continuous chain. Compare: we can say that X is tall whereas Y is not, and yet still say that Z is even taller than X.

A is psychologically continuous with B to a strength of 30 units in mental respect X, and B is continuous with C to a strength of 20 units in the same respect. We can affirm that if A is continuous with B, and B with C, then A is continuous with C, while denying that A is continuous with C to the same strength as A is with B. Continuity is transitive, but its strength isn't. I suggest that A is continuous with C to a strength of 20 units: the weakest point in the overlapping chain of connections.

This seems intuitive. Suppose a deep commitment to honesty runs down a family line. A holds it strongly and transmits it to her child B, who holds it just as strongly and transmits it to C, and so on, generation after generation, each passing it on with full conviction. Suppose this continues unbroken for ninety-nine generations, from A in G_1 to Z in G_{99} . It seems to me that A and Z are strongly continuous with respect to this commitment, even across the ninety-eight generations between them. For the overlapping chain held strongly at every step. Now vary the case. Suppose that at some point—say G_{50} —the commitment faltered: that generation held it only weakly and half-heartedly, even though the chain continued unbroken thereafter. Here, I suspect, A and Z are continuous with respect to their commitment to honesty only to the degree it held at G_{50} , the weakest link in the chain.

Why? For continuity, unlike connectedness, is not a matter of *how much* qualitatively similar mental content is transmitted, but of whether a chain of such transmissions holds *continuously* without gaps. How weakly continuity held just is *how near* the chain came to breaking; how strongly it held just is *how far* it remained from doing so. The chain can then only be said to be as strong as its weakest link. In this sense, intergenerational continuity is like water flowing through a series of connected pipes of differing widths. The flow that gets through is neither the sum of the widths nor a quantity multiplied down at each junction, but the flow the narrowest section allows.

The Continuity Strategy faces several challenges. How can we be continuous with far future people who may be very different from us? Will

the degree of continuity and connectedness remain significant over the long run, or will it eventually shrink to nothing? Can ordinary people reasonably hope for quasi-survival, or is it available only to the especially influential? How should we aggregate what matters across different lives? I will turn to these problems shortly. But let me first show how the view explains and supports many of our intuitions.

5. The Merits of the Continuity Strategy

One intuition we typically have when the object of our prudential concern is preserved without identity—as in *Fission*—is that our death seems less bad, less worrying. Consider how the badness of death varies depending on whether intergenerational psychological continuity is preserved. Imagine lying on your deathbed with a device that lets you foresee the future of humanity. And now contrast two scenarios:

Quasi-Surviving Death. You die discovering that every thought, belief, memory, emotion, and value you care about will continue through one generation after another in a never-ending chain.

Death Without Quasi-Survival. You die discovering that every thought, belief, memory, emotion, and value you care about will die with you in that very moment; that they will not be carried on by anyone else after you.

Dying with continuers seems less bad for you than dying without them. In *Quasi-Surviving Death* much of what matters is carried on through the lives of your continuers, and that seems deeply relieving. In *Death Without Quasi-Survival* all that matters is lost, and that seems deeply disturbing. Our concern for intergenerational continuity captures these intuitions cleanly (and if *Quasi-Surviving Death* still seems not quite as good as ordinary survival, that is because connectedness is lost either way).

One might object that *Quasi-Surviving Death* seems less bad simply because the continuation of the beliefs, memories, emotions, and values we care about matters in an *impersonal* sense. But suppose that, as a violinist, I learn that someone else's chess-playing style will be carried on by future generations of chess players. Even if I take violin and chess to be of roughly equal impartial value, only the continuation of my own violin-playing style would seem to make my death less bad.

Quasi-Surviving Death, then, matters in a *personal* sense. It is good *for* us. Perhaps attaining a high degree of intergenerational continuity can amount to a personal achievement—a standard example of a welfare good. But it doesn't seem necessary to characterise *Quasi-Surviving Death* as an achievement in order to appreciate it. There may be nothing extraordinary about the intergenerational continuities preserved. They might be ordinary memories passed down from a grandmother to her grandchildren, such as the memory of a beautiful afternoon spent together. The survival of such a memory would hardly count as an achievement, but could matter a great deal to the grandmother.

Perhaps, then, this goodness stems from desire-fulfilment.¹⁷⁵ We desire to have a legacy, to be remembered, to transmit our projects, ideas, and values into the future. And fulfilling these preferences may be good for us. But to say that we have a natural aversion to oblivion, to being forgotten by our successors, is simply to restate what demands an explanation. The Continuity Strategy explains *why* some of those preferences that extend beyond our personal extinction are justified—what makes them *rational*, the sort of preferences it is sensible for us to have.

Perhaps some preferences stand in no need of justification. I just happen to like chocolate. But even that preference can be justified by the fact

¹⁷⁵ Note, however, that there can be cases where desire and what matters come apart, as in the cases I considered in Chapter 3 (Section 5). One might take such cases to show either that intergenerational continuity is necessary for posthumous benefit, or that desire-fulfilment and intergenerational continuity are each sufficient for it.

that eating chocolate feels good. Our preference for quasi-survival cannot be justified in that way, and it is an especially odd one to have. For we will no longer exist when we quasi-survive. Therefore it does seem to stand in need of justification.

The Continuity Strategy gives us a criterion for determining which desires to leave a given legacy are rationally justified on egoistic grounds and which are not. Some very common ways in which people tend to think about their legacies may in fact be irrational. Kings and nobles throughout history, for example, have often pursued lasting fame for their names. But a name alone doesn't preserve anything of one's mental life. Stripped of the memories, thoughts, and affections we associate with it, a name is nothing but a word.

If it is not always appropriate to utilise the resources of our rational egoistic concern in the pursuit of posthumous states of affairs, then we should ask which legacies we are rationally justified in pursuing. It is tempting to think that these are those legacies intended to carry forward our ideas, values, beliefs, or other aspects of our mental life. That is partly right. But imagine using again the foreseeing device, this time to discover the following:

Quasi-Surviving Death 2. You die discovering that your laziness, your memories of being brutalised in war, and all your false beliefs will continue through one generation after another in a never-ending chain, inspired by your example.

Continuities of this kind don't make your death less bad for you. The desirability of quasi-survival depends on the *quality* of the relevant intergenerational continuities.¹⁷⁶ Some continuities, indeed, are better than others for the one who is continued. The best carry forward things that are true, good, or beautiful; the worst, things that are false, bad, or ugly.

I find *Quasi-Surviving Death 2* even worse than dying without continuers. Just as ordinary survival is good only if we have some goods to

¹⁷⁶ See Chapter 3 (Section 3.2) for a discussion of the quality of psychological continuities.

anticipate, quasi-survival is good only if our continuers have some goods to anticipate. In *Quasi-Surviving Death 2*, however, the relation we bear to our continuers contains only bad things. On reflection, this brings out the anticipatory concern we have for the wellbeing of our continuers. My laziness, wartime traumas, and false beliefs make my life worse than it would otherwise be. They harm me. I do not want future people to be my continuers with respect to such self-destructive traits. They would be harmed too. And since they are *my* continuers, this would be prudentially relevantly like me reliving some bad moments of my life. I would rather die and be forgotten.

Quasi-survival, then, like survival, is in itself neither good nor bad. Whether quasi-survival is good depends on how good the lives of one's continuers are for them, which in turn partly depends on how good or bad the continuities themselves are for them (their quality). If it's all bad, we might wish not to have such continuers, and we would have reason to prevent ourselves from having them, in the same way that, knowing a bad life lay ahead of us, we might be justified in cutting it short. We can think of this as a form of euthanasia with respect to quasi-survival.

One might find the whole idea of an anticipatory concern for the wellbeing of one's continuers disconcerting.¹⁷⁷ But its intuitive plausibility depends on the degree of interpersonal continuity and connectedness. The Continuity Strategy doesn't imply that if I tell someone a random story, who tells it to the next person, and so on for a hundred iterations, I should have significant prudential concern for whether the final person cuts their finger. It implies, rather, that if there will be future people psychologically related to me in numerous and significant respects, then I should have significant prudential concern for their wellbeing. This is more intuitive. Consider:

Your Continuers. You care about a set of beliefs, values, and activities that exist only in your town, nowhere else. On your

¹⁷⁷ See, for example, McMahan's analysis of the *Suicide Mission* case (2002, 57).

deathbed, you are given a forecasting device that lets you see the far future. You see that there will be one surviving line of continuers, who will carry on all your traditions and care about everything you cared about. But you also see that they face 100 years of suffering, unless you now press a button that would divert an equal amount of suffering onto unrelated strangers.

Many of us would feel considerable pressure to press the button, even if we conceded that there were also strong moral reasons against doing so.¹⁷⁸ If we were choosing entirely on the basis of what matters to us prudentially, I suspect many of us would press the button without hesitation. This is not simply due to a greater degree of sympathy: we might be willing to pay a high price, or make other important sacrifices, to divert the suffering onto someone else's line of continuers instead of our own. So we do seem to have a special concern for the interests of our continuers.¹⁷⁹

The thought experiments I have offered are intended to elicit intuitions about our actual psychological relations with future generations. *Quasi-Surviving Death*, for example, doesn't show only that our death may be less bad than we think, but also that the death of humanity may be, in one respect, worse—everything that matters would be lost with it. Here again, the Continuity Strategy captures an intuition we have. That concerning the special, distinctive badness of human extinction.

Consider Scheffler's *Doomsday Scenario* (2013). Imagine learning that after our deaths there will be no future generations and that humankind will soon go extinct. Not only would we react with profound distress, but, in Scheffler's words,

¹⁷⁸ There may be moral reasons to flip a coin, or not to contribute to causing a bad outcome.

¹⁷⁹ For a similar intuition, see McMahan's case of *Multiple Replication*, in which extortionists threaten to create several psychological replicas of you and then torture and kill them unless you transfer all your wealth. McMahan acknowledges that 'one would probably be willing to pay rather more to prevent the torture of replicas of oneself than one would to prevent a comparable fate for an equivalent number of unknown [people]' (2002, 57).

People would lose confidence in the value of many sorts of activities, would cease to see reason to engage in many familiar sorts of pursuits, and would become emotionally detached from many of those activities and pursuits.

(2013, 44)

We might lose much of the motivation driving scientific cancer research, political campaigns, and contributions to the philosophical literature. The significance of these efforts seems to depend not only on the value of the activities themselves, but also on the existence of future generations who will continue or benefit from them. But why is the continuation of humanity so important to us?

To say that we care about the continuation of humanity because we care about projects that crucially extend into the future seems circular. Why do we care about such projects in the first place? Of course, the potentially vast number of lives that might exist in the future demands moral consideration. Without a future for humanity, some of our activities would become pointless because they were intended to benefit posterity. This is part of the answer. But it's certainly not the whole story. We don't worry about the continuation of humanity *just* because we desire to promote goodness and welfare. Consider:

The Good Doomsday Scenario. You learn that after your death there will be no future generations of human beings. Instead, an alien civilisation will spontaneously arise on a distant planet 1,000 years after human extinction. They will be very different from humanity but live comparably good lives.

This future seems roughly as good as the continuation of humanity from an impersonal moral point of view. You might feel some relief in comparison to Scheffler's original *Doomsday Scenario*. But again, once you fully realise that humanity will be replaced by a distant alien species that has nothing to do with

us, you are still likely to react with distress. Your dearest projects turn out to be meaningless. All the thoughts you wrote, the artworks you painted, and the scientific discoveries you made won't last for long. None of these aliens will appreciate your efforts. Nor will they experience your feelings, values, and sensibilities, which will be forgotten, eradicated from Earth. If you would rather have a personal connection with the future, then your concern for your successors isn't wholly reducible to the *impersonal* moral concern for the quality of their lives. Interestingly, and perhaps more controversially, it isn't reducible to a *species-relative* concern for human beings either. Consider:

The Species-Relative Doomsday Scenario. You learn that after your death humans will not survive for long as a species. We will be gradually replaced by cohorts of sentient AIs. But since we created them in our image, they will continue most of our values, beliefs, and traditions, carrying on our mental life just as much as human continuers would have done.

Many of us, on reflection, would find our relation to this future roughly as good as our relation to a future full of members of our own species, all else being equal. I doubt that the last generation of human beings would have reason to experience widespread despair, apathy, or a pervasive loss of conviction about the value of their own activities and traditions, which would all continue to flourish and develop as before. We thus have some intuitive evidence to think that what we care about is not the continuation of human genes, but of human-like forms of mentality, which may be only contingently related to our biological hardware. And if the future-directed relations that matter to us are not restricted to those with our biological descendants, we need a *species-neutral* explanation of the special concern we typically reserve for posterity. The Continuity Strategy provides this explanation, in that it

grounds our concern in the persistence of intergenerational psychological continuity and connectedness.¹⁸⁰

Let us take stock. The Continuity Strategy has its merits. It explains why our death seems less bad in *Quasi-Surviving Death*, why we have a special concern for the wellbeing of *Your Continuers*, and why we are so distinctively distressed by the prospect of the *Doomsday Scenario*. It is a unifying theory that explains a wide range of attitudes within a single framework. As such, it deserves serious consideration.

6. Problems for the Continuity Strategy

I now consider three challenges to the Continuity Strategy: the dissimilarity of future humanity, the gradual decay of continuity and connectedness over time, and a problem of aggregation.

6.1 The Dissimilarity Problem

Our successors may be different from us in ways we cannot even conceive. We might relate to future generations as distantly as Palaeolithic humans are related to us—as people who couldn’t have imagined twenty-first-century beliefs and values. In that case, connectedness would be broken. But continuity need not be. What matters for continuity is that each point in the intergenerational chain is properly connected to the adjacent points, not that G_1 and G_{99} stand in a strong relation of similarity. Imagine using the foreseeing device again and discovering the following:

¹⁸⁰ I’m not alone in noticing an analogy between our concern about the *Doomsday Scenario* and our ordinary prudential concern. Altshuler (2017, 2021) argues that the preservation of humanity allows for a genuine form of personal survival, and that our special concern for future generations is grounded precisely in the concern for our own continued existence. I worry that this route to immortality, as Altshuler understands it, violates the transitivity of identity. But we are in substantial agreement if we move the question from personal identity to what matters in survival.

Dissimilar Humanity. The far future of humanity turns out to be vastly dissimilar from you as you are today. However, you remain psychologically continuous with many future continuers to a significant degree, since the evolution of humanity has been gradual, uninterrupted, and rich in psychological connections from generation to generation over millions of years.

In *Dissimilar Humanity*, G_1 is well connected to G_2 , G_2 to G_3 , and so forth. But G_1 doesn't stand in any relevant similarity relation to G_{99} . Despite this, I suspect that many of us would care about *Dissimilar Humanity* much more than an equally flourishing but wholly unrelated civilisation on a distant planet. There is a significant difference between the persistence of *my own* line of successors and some unrelated civilisation of intelligent, sensitive beings. I have more reason to care about a humanity whose history is causally linked to my choices, my actions, my life. I care about the causal chain between my successors and me to the extent that, were I to lose it, and to know that everything I had done in my life would not affect future lives, I would be deeply distressed.

Continuity is also consistent with radical *intrapersonal* qualitative change. Typically, my future self resembles my present self, but deep psychological continuities persist even when I deliberately control and update my beliefs, desires, and values. For instance, when I want to improve myself. The case of improvement illustrates how a high degree of overlapping chains of connections does not imply fixity of character. A life marked by extensive change, where our character is held together by transformative strands of desire and intentionality, can be more deeply and richly unified than a life whose character remains static. There are more bridges, more connections at each point in time. The same is true of our descendants. We may desire a better future and deliberately implement plans to improve the next generations. If

successful, we might establish richer and more numerous continuities precisely by intentionally transforming the character of posterity.

Take Copernicus, for example. True, his heliocentric model is now immortal, and he worked on it for about twenty-five years. But how much of what mattered to Copernicus is present today, or has been preserved by the generations that followed him? It's not *his* theory that survived, but only some concepts from it. While we now accept the sun is at the centre of our solar system, our beliefs about astronomy have evolved far beyond Copernicus's. His theory has *never* been the standard in exactly the form in which he proposed it: the next generation debated Kepler's theory, which was different from Copernicus's, and the generation after Kepler debated Newton's theory, which was different from Copernicus's and Kepler's, and so on. Nonetheless, a significant degree of what mattered to Copernicus may have been preserved through this continuous generational process of improvement. After all, wouldn't Copernicus have appreciated the way Newton used his ideas to investigate new questions even more than the way a historian might carefully explain those exact ideas to some students?

Maybe this is too quick. Continuity can accommodate qualitative diversity, but isn't there a limit? Suppose that what matters to me is to *paint everything black*. It may seem strange to think that what matters to me is preserved if, after gradual and continuous changes over generations, my descendants come to *paint everything white*. Does the Continuity Strategy entail that even members of future generations who consciously work to destroy *all* my values and projects can be continuers worthy of my prudential care?

Perhaps this case counts in favour of the Similarity Strategy I briefly outlined earlier. Or perhaps we would still care about our continuers even if they were to *paint everything white*, much like a parent would still care about their children even if they were to gradually transform into something

unrecognisably evil. We would be very upset, disappointed. But this seems the opposite of indifference.

6.2 The Decay Problem

On the Continuity Strategy, our special concern for future people is a function of the degree of intergenerational continuity and connectedness. But if I shared with them only 0.0001% of the psychological relations that I normally share with my future self, I wouldn't be justified in having much special concern for the far future. We may then have a problem. For it's hard to rule out the possibility of a gradual but relentless decay of these psychological relations over generations. Suppose that (i) only half of the members of each generation have significant connections to some members of the next; that (ii) at each intergenerational step the continuities and direct connections weaken, if only slightly; and that (iii) the future is very long. Then, for each succeeding generation, the degree of prudential concern would decline as the number and closeness of the bonds linking present and future people decline. It follows that intergenerational continuity and connectedness should decay until they drop to, or asymptotically approach, zero.

However, things are messier and more complex than this picture suggests. For some individuals there has been no rapid decay of connectedness and continuity. Quite the opposite: they may have experienced something closer to an exponential increase. The beliefs, values, and virtues of figures such as Plato, Dante, and Shakespeare have contributed to forming the Western consciousness, generating overlapping chains of psychological connections across billions of future people—future, that is, from their temporal perspective. Given that Plato's mental life is continued through both the untold number of connections directly established by his books and their

ripple-effects on other minds, it's not unreasonable to think that much of what mattered to him may have been preserved to this day.¹⁸¹

Of course, the fact that certain influential figures enjoy significant continuity across generations doesn't imply that ordinary people have reason to be prudentially concerned for the future of humanity. The passing-on process may be disproportionately achieved by very few individuals.

I find this thesis plausible—perhaps quasi-survival is a privilege for the elite. But I don't find it trivially true. From the point of view of the history of humanity, ordinary people may partake in significant intergenerational continuities they are unaware of. People such as Copernicus, Van Gogh, and Thoreau died without knowing how their work would shape the lives of the subsequent generations. It is possible that the vast majority of those who changed the course of history have never received any kind of acknowledgement, like perhaps Aristotle's drinking companions. Their names may have been forgotten, but that doesn't prevent others from carrying on their mental life.

And even if we are not friends of the next Aristotle, traces of our mental life may be influential in other ways. Small contributions by ordinary people to large-scale movements or traditions could well establish rich continuities with the far future. Traditions are, after all, multigenerational enterprises 'whose organizing purpose is to preserve what is valued beyond the life span of any single individual.'¹⁸² Such historical movements don't always survive, but our participation in great ones can give us the opportunity to transmit some of our core values into the far future. The Renaissance, for instance, exerts an ongoing influence on humanity's mental life, which never drops to zero thanks to its surviving artefacts. Even legacies confined to one's own family or loved

¹⁸¹ Contemporary philosophical theories differ a great deal from Plato's original theories. But we might interpret the history of philosophy as a gradual process of generational improvement that preserves significant continuities. This need not imply the metaphilosophical claim that there has been objective philosophical progress, but only that, as philosophers, we have continuously striven to improve our theories over the centuries.

¹⁸² Scheffler (2013, 33).

ones may sustain long-lasting continuities unified by a shared allegiance to a set of values. Not all family members will share the same values, of course, and there will be frequent gaps or weak links in the chains of connections. But within each family line, the number of descendants typically grows exponentially over time as family trees branch out, and the number of continuers among them may increase accordingly. It then becomes statistically more plausible that an ordinary parent may have at least a few branches of the family tree along which a particular set of values or beliefs of theirs is preserved without gaps and held strongly at every point in the intergenerational chain.

Moreover, even if each of us were in fact only very weakly psychologically continuous or connected to each future continuer, we might nonetheless have reason to be considerably prudentially concerned for the future of humanity as a whole. The weakening of continuity and connectedness could be compensated for by the increasing number of continuers (I'll return to this point shortly).

I don't take these preliminary suggestions as conclusive arguments. They show only that the decay problem isn't a conclusive argument either. At least for some individuals, the psychological connections and continuities that hold between them and distant future generations seem numerous and significant. And it is unclear whether intergenerational continuity ever drops to zero for most ordinary people.¹⁸³ Given this uncertainty, it may be safer for us to act as if we *will* enjoy a significant degree of intergenerational continuity. This cautious approach would minimise the risk of throwing away the chance to preserve what matters after we die.

¹⁸³ There are of course differences in how confident different people should be about their degree of intergenerational continuity and connectedness. A hermit, for example, seems unlikely to have much influence on the psychological states of future people.

6.3 The Disintegration Objection

In many cases of quasi-survival, our prudential concern should extend to a vast number of continuers. But isn't it better to have a single happy duplicate who is maximally psychologically continuous with you than to have 100,000 happy continuers each very weakly continuous with you? Can what matters survive disintegration at all?

Some theorists find disintegration unacceptable. McMahan, for example, says that a Parfitian understanding of psychological connection is too liberal and 'may lead to an implausible proliferation of psychological connections' (2002, 61). Call this the *Disintegration Objection*: preserving your mentality across multiple lives could never be as good as preserving it within a single life. What matters is an organic unity, where the whole is greater than the sum of its parts; if these parts are scattered across different entities, then the prudential value they had in unification is lost. Or so the objection goes.

Peter Unger (1990, 255–287) formulates this worry in an especially compelling way. In his view, if you were to distribute your prudential concern across a vast number of continuers (he considers fission-cases), you would experience *a loss of focus* in your life. To illustrate, he draws an analogy between an ordinary man with two children and a polygamous sultan with one hundred children. The ordinary man, Unger says, can plausibly take a much more intense special interest in each of his two children than the sultan can in each of his hundred. While this analogy may seem intuitive, I have four concerns about Unger's example, which correspond to four ways of dealing with the Disintegration Objection.

First, the loss of focus in one's life isn't necessarily a normative loss. Just as a sultan cannot properly worry about all his one hundred children, perhaps we cannot properly worry about all the continuers we will have in the far future. But this does not mean that we *shouldn't* worry about them; it only

means that the cognitive demand would be unreasonable given our limited attentional and evaluative capacities. If the sultan had more time and energy, he could, and plausibly should, devote himself to each of his children as an ordinary father does. Likewise, with the support of enhanced cognitive capacities, it could be perfectly rational for us to be concerned for the wellbeing of all those individuals who carry on our mental life into the future.

Second, the loss of focus in one's life doesn't seem to be the loss of everything that matters. The sultan still cares a great deal about his children and, similarly, we may still care a great deal about our continuers even if we accepted that quasi-survival can never be as good as ordinary survival.

Third, the loss of focus might not be a prudential loss at all. Sure, the sultan isn't concerned for each of his one hundred children as an ordinary father is concerned for each of his two children. But collectively, the sultan might care about his offspring as much as the ordinary father cares about his own. This seems like a fairer comparison. The sultan's special concern is not diminished; it's just distributed differently. He can still say things such as, 'My children are my life.' Likewise, if future people are connected to me much more weakly than I am connected to my own future self, but there are many, many more of them, then most or even all of what matters to me could survive my lifetime despite its disintegration. Perhaps this is the case for martyrs who refuse to recant their faith and thereby inspire many others for a long time, or for dissidents in totalitarian states who spread their ideas at the risk of their lives. How much of what matters is preserved when they die? Is it possible that dying is, in fact, the prudentially rational choice, and that by dying they can preserve more of what matters to them than if they had lived? Whether this could ever be the case depends on how we should aggregate what matters across different lives. Perhaps having many feeble continuers could never be as good as ordinary survival, however great their number. Or perhaps there is, after all, some threshold number of continuers that would suffice, however tenuous their continuities. Answering this would require developing a theory

of interpersonal aggregation—a notoriously difficult task, far beyond the scope of this thesis.¹⁸⁴ So I will simply invite the reader to take the Continuity Strategy and modify it as they wish, according to their favourite theory of aggregation.

Fourth, the loss of focus in one’s life could actually be a benefit. True, the sultan misses the prudentially valuable opportunity to concentrate his care on very few people and love them more intensely than anything else. But if so, then it’s also true that an ordinary father misses the prudentially valuable opportunity to widen his love to one hundred children. That’s a loss too. If we think of a person maximally and exclusively focused on their own life, it’s easy to see that there is something missing in their inability to expand their egoistic perspective. A wider focus seems better for us. Other people become closer. We are less on our own. And if enough of what we care about in survival will persist through intergenerational bonds, then we can worry less about our own personal extinction. It may not be immortality, but it could be the next best thing: a long future full of what matters to us.

7. Conclusion

I began with the question ‘Will *I* live on through future generations?’ and argued that, to avoid insurmountable problems, we must frame the question differently. What we need to ask is: ‘Will *what matters to me in my ordinary survival* live on through future generations?’

I then put forward the Continuity Strategy for quasi-survival, according to which a legacy of what matters will be passed on through the generations to come, proportionally to the degree of intergenerational

¹⁸⁴ Gustafsson and Kosonen (2022, 2025) offer one of the first investigations of this sort of interpersonal prudential axiology. Interestingly, they argue that we should accept a version of the Prudential Total View, or some similar additive principle, which, combined with the Continuity Strategy, would imply that having many continuers could be as good as ordinary survival, or even better.

psychological continuity and connectedness holding between us and our successors.

The degree of quasi-survival, and the prudential care we should accordingly have for future generations, turn on a multifaceted set of factors: the strength, number, and quality of the psychological connections that will causally influence the distant future, the right theory of aggregation, and the existential risks humanity will face. There is, nonetheless, reason to hope that a significant part of what matters to us will survive through our continuers. This means we have an additional reason to care about future generations—a reason of quasi-survival.

Concluding Remarks

This integrated thesis has addressed a wide range of questions about mind and value across four independent chapters. It is now worth drawing these threads together: summarising the main results, showing how they stand in relation to one another, and asking what broader philosophical picture emerges from them.

In Chapter 1, I argued for *Phenomenal Necessitarianism*: the view that *only* conscious beings can be welfare subjects, have lives that can go well or badly for them, accrue welfare goods or bads, and the like. I argued that we have powerful introspective evidence in support of this belief, and that the candidate defeaters are either weak or can be accommodated by the necessitarian requirement at no significant extra cost. It would therefore be irrational to give up this belief in favour of a different theoretical package, since the desideratum of accommodating this sort of introspective data overrides any non-significant differences in theoretical cost between competing theories. All things considered, then, we should be necessitarians.

In Chapter 2, I argued for a more radical thesis, and one likely to generate more disagreement, even among sentientists already committed to Necessitarianism. I maintained that consciousness is necessary for welfare subjectivity, but not sufficient. The kind of mind that makes a being a welfare subject is *affective consciousness*: the capacity for experiences that feel good or bad, such as pleasure, pain, and emotion. This is because, as I argued, a welfare subject is best understood as a being capable of accruing welfare goods or bads, and each candidate welfare good is itself best understood as having an affective component. Without such a component, these goods lose their original intuitive plausibility, and with it the very reason we took them to be candidate welfare goods in the first place.

Chapters 1 and 2 together argued that *all* and *only* affectively conscious beings are welfare subjects. They defended this view against a wide range of arguments for the claim that some non-conscious beings, or some affectless conscious beings, are welfare subjects. I called the resulting view *Affective Sentientism*, a form of Sentientism that neither restricts welfare subjectivity to beings like us nor extends it to every system that behaves intelligently, pursues goals, or represents the world. Moreover, Chapter 2 argued that, while affect grounds welfare subjectivity, welfare subjectivity grounds moral status. Many traditions ground moral status elsewhere—in rational agency, autonomy, species membership, or relations—but the arguments in Chapter 2 ruled out these options both as necessary and as sufficient for it.

Many philosophers have argued for an expansion of our *moral circle*, that is, the circle of beings whose interests should be taken into account in our moral deliberations.¹⁸⁵ I agree that we do not yet treat many beings, especially many animals, as having morally significant interests. They deserve immeasurably better treatment. But these chapters also argued for a certain restriction on the class of welfare subjects. We should be careful about which beings we count among them, not least because over-attribution risks misdirecting our limited resources. On the view defended here, welfare value is a rare and precious thing in the universe: rare among the things that exist, and precious wherever it occurs. It is possessed only by certain kinds of beings in virtue of how they feel, rather than in virtue of how they behave, what they look like, what they think, or how we feel about them. So yes, we should expand our moral circle. But we should expand it on the right grounds, and not as far as some theorists think.

In the second part of the thesis, I defended a more revisionary claim. Even if only a certain kind of mind makes one a bearer of welfare goods or bads, it is relatively easy to connect one mind to another in a way that gives one a reason to care about the welfare of others in the distinctively egoistic,

¹⁸⁵ Singer (1977), Sebo (2025).

prudential way one should care about one's own. All that is needed, as Parfit famously argued, is suitable psychological similarity and some causal connection between one's mental states and those of others.

In Chapter 3, I argued that even if only conscious beings matter morally, there is still an important sense in which beings who were once conscious can be harmed or benefited after death. This is because something of their mental life, namely certain kinds of psychological connections, can survive in other minds. The resulting theory has two main virtues. The first is that it locates welfare value only in concrete, conscious individuals. Nothing needs to accrue to the dead: my extended wellbeing is nothing over and above the wellbeing of those who continue my mental life, and it is affected at the time their wellbeing is affected. This is why the view requires no backward determination of welfare value, and why it is, so far as I know, the only theory of posthumous harm in principle compatible with experientialism about welfare. The second is that the theory is, I argued, extensionally adequate. It explains why damage to a posthumous reputation is bad for the deceased, why we have reasons of remembrance, and why completing a dead person's unfinished projects can benefit her without presupposing that achievement or desire-fulfilment is a welfare good. It even grounds reparations to the dead victims of historical injustice, among other things.

In Chapter 4, I extended this same argument from the past to the future. I argued that because of these surviving psychological connections, a significant degree of what we care about in our own continued existence can be expected to live on through future generations, and that we therefore have prudential reasons to care about them. I defended the view that a significant degree of continuity and connectedness can be expected to persist through the various changes and disruptions that might occur throughout the future history of humanity. I argued that this account offers a unified and principled explanation of a range of attitudes we already have: why dying seems less bad when we know that what we care about will be carried forward, why we have

a distinctive concern for the wellbeing of those who will continue our mental life, and why the prospect of human extinction distresses us in the particular way it does. It also has revisionary consequences. Many believe that we have moral reasons to care about future generations. If I am right, we have prudential reasons too, and therefore stronger reasons overall to care about them. Our own deaths may be less bad than we thought, while the extinction of humanity could be worse.

The general picture that emerges from these four chapters is one on which a certain kind of conscious life is the locus of welfare and moral status. But the prudential significance of a conscious life is not confined within the boundaries of the individual who lives it. Through psychological continuity and connectedness, conscious lives can be linked across persons and generations, across space and time, in ways that give us egoistic reasons to care about the lives of others. Only conscious lives matter, but the walls between them are thinner than they seem.

References

- Altshuler, R. (2017). 'Bootstrapping the Afterlife.' *Journal of Moral Philosophy*, 14 (2): 201–216.
- Altshuler, R. (2021). 'The Significance of Future Generations.' In M. Cholbi & T. Timmerman (eds.), *Exploring the Philosophy of Death and Dying: Classic and Contemporary Perspectives*. New York: Routledge, 191–199.
- Anderson, S. W. et al. (1999). 'Impairment of Social and Moral Behavior Related to Early Damage in Human Prefrontal Cortex.' *Nature Neuroscience*, 2: 1032–1037.
- Aristotle. (1984). *Metaphysics*. W. D. Ross (trans.). In J. Barnes (ed.), *The Complete Works of Aristotle: The Revised Oxford Translation*, vol. 2. Princeton: Princeton University Press.
- Aristotle. (1953). *Nicomachean Ethics*. J. Thomson (ed.). Baltimore, MD: Penguin Books.
- Arneson, R. (1999). 'Human Flourishing Versus Desire Satisfaction.' *Social Philosophy and Policy*, 16 (1): 113–142.
- Baker, C. (2023). 'Buddhism and Utilitarianism.' In R. Y. Chappell, D. Meissner & W. MacAskill (eds.), *An Introduction to Utilitarianism*. <https://www.utilitarianism.net/guest-essays/buddhism-and-utilitarianism>.
- Baker, C. (2024). 'Is Buddhism Without Rebirth 'Nihilism with a Happy Face'?' *Analysis*, 84 (4): 701–710.
- Basl, J. (2019). *The Death of the Ethic of Life*. New York: Oxford University Press.
- Bechara, A. et al. (2000). 'Characterization of the Decision-Making Deficit of Patients with Ventromedial Prefrontal Cortex Lesions.' *Brain*, 123: 2189–2202.
- Belliotti, R. A. (2011). *Posthumous Harm: Why the Dead Are Still Vulnerable*. Lanham, MD: Lexington Books.
- Bentham, J. (1907). *An Introduction to the Principles of Morals and Legislation*. Oxford: Clarendon Press [1780].

- Berridge, K. & T. Robinson. (1998). 'What Is the Role of Dopamine in Reward: Hedonic Impact, Reward Learning, or Incentive Salience?' *Brain Research Reviews*, 28: 309–369.
- Birch, J. (2022). 'Materialism and the Moral Status of Animals.' *The Philosophical Quarterly*, 72 (4): 795–815.
- Birch, J. (2024). *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford: Oxford University Press.
- Birch, J. et al. (2021). *Review of the Evidence of Sentience in Cephalopod Molluscs and Decapod Crustaceans*. London: Department for Environment, Food and Rural Affairs.
- Block, N. (1995). 'On a Confusion about a Function of Consciousness.' *Behavioral and Brain Sciences*, 18 (2): 227–247.
- Boonin, D. (2019). *Dead Wrong: The Ethics of Posthumous Harm*. Oxford: Oxford University Press.
- Bradford, G. (2015). *Achievement*. Oxford: Oxford University Press.
- Bradford, G. (2022). 'Consciousness and Welfare Subjectivity.' *Noûs*, 57 (4): 905–921.
- Bradley, B. (2009). *Well-Being and Death*. New York: Oxford University Press.
- Bramble, B. (2016). 'A New Defense of Hedonism about Well-Being.' *Ergo: An Open Access Journal of Philosophy*, 3: 85–112.
- Brandt, R. (1979). *A Theory of the Good and the Right*. Oxford: Clarendon Press.
- Brink, D. (1990). 'Rational Egoism, Self, and Others.' In O. Flanagan & A. Rorty (eds.), *Identity, Character, and Morality*. Cambridge, MA: MIT Press.
- Brink, D. (1997a). 'Rational Egoism and the Separateness of Persons.' In J. Dancy (ed.), *Reading Parfit*. Oxford: Blackwell, 96–134.
- Brink, D. (1997b). 'Self-Love and Altruism.' *Social Philosophy and Policy*, 14 (1): 122–157.
- Brink, D. (2021). 'Special Concern and Personal Identity.' In J. McMahan, T. Campbell, J. Goodrich & K. Ramakrishnan (eds.), *Principles and Persons: The Legacy of Derek Parfit*. Oxford: Oxford University Press, 15–38.

- Campbell, S. (2005). 'Is Causation Necessary for What Matters in Survival?' *Philosophical Studies*, 126 (3): 375–396.
- Carruthers, P. (1999). 'Sympathy and Subjectivity.' *Australasian Journal of Philosophy*, 77 (4): 465–482.
- Cecchinato, M. (2023). 'Double-Standard Moralism: Why We Can Be More Permissive Within Our Imagination.' *British Journal of Aesthetics*, 64 (1): 67–87.
- Cecchinato, M. (forthcoming). 'The Mind that Matters.' *The Philosophical Quarterly*.
- Chalmers, D. (2022). *Reality+: Virtual Worlds and the Problems of Philosophy*. New York: Penguin Books.
- Chalmers, D. (forthcoming). 'Sentience and Moral Status.' In G. Lee & A. Pautz (eds.), *The Importance of Being Conscious*. Oxford: Oxford University Press.
- Chudnoff, E. (2013). *Intuition*. New York: Oxford University Press.
- Crisp, R. (2001). 'Well-Being.' In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition).
- Crisp, R. (2006). *Reasons and the Good*. Oxford: Oxford University Press.
- Cutter, B. (2017). 'The Metaphysical Implications of the Moral Significance of Consciousness.' *Philosophical Perspectives*, 31 (1): 103–130.
- Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Putnam.
- DeGrazia, D. (1996). *Taking Animals Seriously: Mental Life and Moral Status*. New York: Cambridge University Press.
- de Lazari-Radek, K. & P. Singer. (2014). *The Point of View of the Universe: Sidgwick and Contemporary Ethics*. Oxford: Oxford University Press.
- Delon, N. (2024). 'Strangers to Ourselves: A Nietzschean Challenge to the Badness of Suffering.' *Inquiry: An Interdisciplinary Journal of Philosophy*, 67 (9): 3600–3629.

- Dennett, D. (1991). *Consciousness Explained*. Boston: Little, Brown and Company.
- Dimitrov, M. (2024). 'Direct Acquaintance with Intrinsic Value.' *The Philosophical Quarterly*, 75 (2): 428–449.
- Dorsey, D. (2013). 'Desire-Satisfaction and Welfare as Temporal.' *Ethical Theory and Moral Practice*, 16 (1): 151–171.
- Dorsey, D. (2022). 'Friendship and the Wishes of the Dead.' *Legal Theory*, 28 (2): 124–145.
- Elliot, R. (1991). 'Personal Identity and the Causal Continuity Requirement.' *The Philosophical Quarterly*, 41: 55–75.
- Enoch, D. (2016). 'What's Wrong with Paternalism: Autonomy, Belief, and Action.' *Proceedings of the Aristotelian Society*, 116 (1): 21–48.
- Fanciullo, J. (2025a). 'Alienation, Engagement, and Welfare.' *The Philosophical Quarterly*, 75 (1): 40–60.
- Fanciullo, J. (2025b). 'Are Current AI Systems Capable of Well-Being?' *Asian Journal of Philosophy*, 4 (1): 1–10.
- Feinberg, J. (1974). 'The Rights of Animals and Unborn Generations.' In W. T. Blackstone (ed.), *Philosophy and Environmental Crisis*. Athens, GA: The University of Georgia Press, 43–68.
- Feinberg, J. (1977). 'Harm and Self-Interest.' In P. M. S. Hacker & J. Raz (eds.), *Law, Morality, and Society: Essays in Honour of H. L. A. Hart*. Oxford: Clarendon Press, 285–398.
- Feinberg, J. (1984). *Harm to Others*. New York: Oxford University Press.
- Feldman, F. (1991). 'Some Puzzles About the Evil of Death.' *The Philosophical Review*, 100 (2): 205–227.
- Finnis, J. (1979). *Natural Law and Natural Rights*. New York: Oxford University Press.
- Fodor, J. (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford: Oxford University Press.

- Frankfurt, H. G. (2004). *The Reasons of Love*. Princeton: Princeton University Press.
- Frankl, V. E. (1959). *Man's Search for Meaning*. Boston: Beacon Press.
- Gardner, M. (2022). 'David Boonin, *Dead Wrong: The Ethics of Posthumous Harm* [Book Review].' *Ethics*, 132 (4): 886–889.
- Georgalis, N. (2006). *The Primacy of the Subjective: Foundations for a Unified Theory of Mind and Language*. Cambridge, MA: MIT Press.
- Goldstein, S. & C. D. Kirk-Giannini. (2025). 'AI Wellbeing.' *Asian Journal of Philosophy*, 4 (1): 1–22.
- Grahek, N. (2007). *Feeling Pain and Being in Pain*, 2nd edn. Cambridge, MA: MIT Press.
- Griffin, J. (1986). *Well-Being: Its Meaning, Measurement, and Moral Importance*. Oxford: Clarendon Press.
- Grover, D. (1989). 'Posthumous Harm.' *The Philosophical Quarterly*, 39 (156): 334–353.
- Gustafsson, J. E. & P. Kosonen. (2022). 'Do Lefty and Righty Matter More Than Lefty Alone?' *Erkenntnis*, 89 (5): 1921–1926.
- Gustafsson, J. E. & P. Kosonen. (2025). 'Prudential Longtermism.' In H. Greaves, J. Barrett & D. Thorstad (eds.), *Essays on Longtermism: Present Action for the Distant Future*. Oxford: Oxford University Press.
- Harman, E. (2003). 'The Potentiality Problem.' *Philosophical Studies*, 114 (1): 173–198.
- Heathwood, C. (2019). 'Which Desires Are Relevant to Well-Being?' *Noûs*, 53 (3): 664–688.
- Huemer, M. (2001). *Skepticism and the Veil of Perception*. Lanham, MD: Rowman & Littlefield.
- Hurka, T. (1993). *Perfectionism*. New York: Oxford University Press.
- Jackson, F. (1982). 'Epiphenomenal Qualia.' *The Philosophical Quarterly*, 32: 127–136.

- Jaworska, A. (2007). 'Caring and Full Moral Standing.' *Ethics*, 117 (3): 460–497.
- Jenkins, J. (2011). 'Dead and Gone.' *Utilitas*, 23 (2): 228–234.
- Jeske, D. (1993). 'Persons, Compensation, and Utilitarianism.' *The Philosophical Review*, 102 (4): 541–575.
- Johansson, J. (2012). 'The Time of Death's Badness.' *Journal of Medicine and Philosophy*, 37 (5): 464–479.
- Johnston, M. (2010). *Surviving Death*. Princeton: Princeton University Press.
- Kagan, S. (2009). 'Well-Being as Enjoying the Good.' *Philosophical Perspectives*, 23 (1): 253–272.
- Kagan, S. (2019). *How to Count Animals, More or Less*. Oxford: Oxford University Press.
- Kahane, G. (2010). 'Feeling Pain for the Very First Time: The Normative Knowledge Argument.' *Philosophy and Phenomenological Research*, 80 (1): 20–49.
- Kamm, F. M. (2007). *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. New York: Oxford University Press.
- Kolak, D. & R. Martin. (1987). 'Personal Identity and Causality: Becoming Unglued.' *American Philosophical Quarterly*, 24 (4): 339–347.
- Korsgaard, C. M. (2018). *Fellow Creatures: Our Obligations to the Other Animals*. Oxford: Oxford University Press.
- Kraut, R. (2018). *The Quality of Life: Aristotle Revised*. Oxford: Oxford University Press.
- Kriegel, U. (2019). 'The Value of Consciousness.' *Analysis*, 79 (3): 503–520.
- Kriegel, U. (forthcoming). 'The Value of Consciousness to the One Who Has It.' In G. Lee & A. Pautz (eds.), *The Importance of Being Conscious*. Oxford: Oxford University Press.
- Lee, A. Y. (2019). 'Is Consciousness Intrinsically Valuable?' *Philosophical Studies*, 176 (3): 655–671.

- Lee, A. Y. (2022). 'Speciesism and Sentientism.' *Journal of Consciousness Studies*, 29 (3–4): 205–228.
- Lee, A. Y. (2023a). 'Degrees of Consciousness.' *Noûs*, 57 (3): 553–575.
- Lee, A. Y. (2023b). 'The Neutrality of Life.' *Australasian Journal of Philosophy*, 101 (3): 685–703.
- Lee, A. Y. (2025). 'Consciousness Makes Things Matter.' *Philosophers' Imprint*, 25 (35): 1–18.
- Lee, A. Y. (forthcoming). 'Metaethical Experientialism.' In G. Lee & A. Pautz (eds.), *The Importance of Being Conscious*. Oxford: Oxford University Press.
- Levenbook, B. B. (1984). 'Harming Someone After His Death.' *Ethics*, 94 (3): 407–419.
- Levy, N. (2014). 'The Value of Consciousness.' *Journal of Consciousness Studies*, 21 (1–2): 127–138.
- Levy, N. (2020). 'It Might Not Matter Very Much Whether Insects Are Conscious.' *Animal Sentience*, 29 (2).
- Levy, N. (2024). 'Consciousness Ain't All That.' *Neuroethics*, 17 (2): 1–14.
- Lewis, D. (1976). 'Survival and Identity.' In A. Rorty (ed.), *The Identities of Persons*. Berkeley: University of California Press, 17–40.
- Lin, E. (2017). 'Asymmetrism about Desire Satisfactionism and Time.' In M. C. Timmons (ed.), *Oxford Studies in Normative Ethics*, vol. 7. Oxford: Oxford University Press, 161–183.
- Lin, E. (2021). 'The Experience Requirement on Well-Being.' *Philosophical Studies*, 178: 867–886.
- Lin, E. (manuscript). 'Welfare Subjects.'
- Lucas, R. & E. Diener. (2008). 'Personality and Subjective Well-Being.' In O. P. John, R. W. Robins & L. A. Pervin (eds.), *Handbook of Personality: Theory and Research*, 3rd edn. New York: The Guilford Press.
- Luper, S. (2004). 'Posthumous Harm.' *American Philosophical Quarterly*, 41 (1): 63–72.

- Marcus, R. B. (1990). 'Some Revisionary Proposals about Belief and Believing.' *Philosophy and Phenomenological Research*, 50: 133–153.
- McMahan, J. (1988). 'Death and the Value of Life.' *Ethics*, 99 (1): 32–61.
- McMahan, J. (2002). *The Ethics of Killing: Problems at the Margins of Life*. New York: Oxford University Press.
- McMahan, J. (2016). 'On 'Modal Personism'.' *Journal of Applied Philosophy*, 33 (1): 26–30.
- Mendola, J. (1990). 'Objective Value and Subjective States.' *Philosophy and Phenomenological Research*, 50 (4): 695–713.
- Mill, J. S. (1998). *Utilitarianism*. R. Crisp (ed.). Oxford: Oxford University Press [1863].
- Mogensen, A. L. (2022). 'The Only Ethical Argument for Positive δ ? Partiality and Pure Time Preference.' *Philosophical Studies*, 179: 2731–2750.
- Mogensen, A. L. (2025a). 'Desire-Fulfilment and Consciousness.' *Philosophical Studies*, 183 (2): 511–529.
- Mogensen, A. L. (2025b). 'Once More, Without Feeling.' *Philosophy and Phenomenological Research*, 111 (1): 343–365.
- Nagel, T. (1970). 'Death.' *Noûs*, 4 (1): 73–80.
- Nozick, R. (1974). *Anarchy, State, and Utopia*. New York: Basic Books.
- Nussbaum, M. C. (2013). 'The Damage of Death: Incomplete Arguments and False Consolations.' In J. S. Taylor (ed.), *The Metaphysics and Ethics of Death: New Essays*. New York: Oxford University Press, 25–43.
- Nussbaum, M. C. (2022). *Justice for Animals: Our Collective Responsibility*. New York: Simon & Schuster.
- Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity*. London: Hachette.
- Parfit, D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.
- Parfit, D. (1993). 'The Indeterminacy of Identity: A Reply to Brueckner.' *Philosophical Studies*, 70 (1): 23–33.

- Parfit, D. (2011). *On What Matters*, vol. 1. Oxford: Oxford University Press.
- Perry, M. (2024). 'Why Sentience Should Be the Only Basis of Moral Status.' *The Journal of Ethics*, 28 (4): 719–741.
- Pitcher, G. (1984). 'The Misfortunes of the Dead.' *American Philosophical Quarterly*, 21 (2): 183–188.
- Portmore, D. (2007). 'Desire Fulfilment and Posthumous Harm.' *American Philosophical Quarterly*, 44 (1): 27–38.
- Powell, R. & I. Mikhalevich. (2020). 'Affective Sentience and Moral Protection.' *Animal Sentience*, 29 (35).
- Pryor, J. (2000). 'The Skeptic and the Dogmatist.' *Noûs*, 34 (4): 517–549.
- Quinn, W. (1993). 'Putting Rationality in Its Place.' In *Morality and Action*. Cambridge: Cambridge University Press, 228–255.
- Railton, P. (1986). 'Facts and Values.' *Philosophical Topics*, 14 (2): 5–31.
- Ramsey, F. P. (1928). 'A Mathematical Theory of Saving.' *The Economic Journal*, 38 (152): 534–559.
- Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.
- Ridge, M. (2003). 'Giving the Dead Their Due.' *Ethics*, 114 (1): 38–59.
- Roelofs, L. (2023). 'Sentientism, Motivation, and Philosophical Vulcans.' *Pacific Philosophical Quarterly*, 104: 301–323.
- Scheffler, S. (2013). *Death and the Afterlife*. N. Kolodny (ed.). Oxford: Oxford University Press.
- Schwitzgebel, E. (2008). 'The Unreliability of Naive Introspection.' *The Philosophical Review*, 117 (2): 245–273.
- Schwitzgebel, E. (2016). 'Phenomenal Consciousness, Defined and Defended as Innocently as I Can Manage.' *Journal of Consciousness Studies*, 23 (11-12): 224–235.
- Sebo, J. (2025). *The Moral Circle*. New York: W. W. Norton & Company.

- Serafini, A. (1990). 'Callahan on Harming the Dead.' *Journal of Philosophical Research*, 15: 329–339.
- Shepherd, J. (2018). *Consciousness and Moral Status*. New York: Routledge.
- Shepherd, J. (2023). 'Non-Human Moral Status: Problems with Phenomenal Consciousness.' *AJOB Neuroscience*, 14 (2): 148–157.
- Shepherd, J. (2024). 'Sentience, Vulcans, and Zombies: The Value of Phenomenal Consciousness.' *AI and Society*, 39 (6): 3005–3015.
- Siderits, M. (1988). 'Ehring on Parfit's Relation R.' *Analysis*, 48 (1): 29–32.
- Sidgwick, H. (1907). *The Methods of Ethics*, 7th edn. London: Macmillan.
- Siewert, C. (1998). *The Significance of Consciousness*. Princeton: Princeton University Press.
- Siewert, C. (manuscript). 'How Subjectivity Makes Lives Matter.'
- Singer, P. (1977). *Animal Liberation: A New Ethics for Our Treatment of Animals*. New York: Avon Books.
- Singer, P. (1993). *Practical Ethics*, 2nd edn. Cambridge: Cambridge University Press.
- Sinhababu, N. (2024). 'The Epistemic Argument for Hedonism.' In S. Chakraborty (ed.), *Human Minds and Cultures*. Switzerland: Springer Nature, 137–158.
- Slote, M. (2009). *Moral Sentimentalism*. New York: Oxford University Press.
- Smith, M. (1987). 'The Humean Theory of Motivation.' *Mind*, 96: 36–61.
- Smithies, D. (2026). 'Hedonic Consciousness and Moral Status.' In U. Kriegel (ed.), *Oxford Studies in Philosophy of Mind*, vol. 5. Oxford: Oxford University Press.
- Sperling, D. (2008). *Posthumous Interests: Legal and Ethical Perspectives*. Cambridge: Cambridge University Press.
- Stalnaker, R. (1984). *Inquiry*. Cambridge, MA: MIT Press.
- Strawson, G. (1994). *Mental Reality*. Cambridge, MA: MIT Press.

- Sumner, L. W. (1996). *Welfare, Happiness, and Ethics*. Oxford: Oxford University Press.
- Thompson, E. (2014). *Waking, Dreaming, Being: Self and Consciousness in Neuroscience, Meditation, and Philosophy*. New York: Columbia University Press.
- Unger, P. (1990). *Identity, Consciousness, and Value*. New York: Oxford University Press.
- van der Deijl, W. (2019). 'Is Pleasure All That Is Good About Experience?' *Philosophical Studies*, 176 (7): 1–19.
- van der Deijl, W. (2021). 'The Sentience Argument for Experientialism about Welfare.' *Philosophical Studies*, 178 (1): 187–208.
- van der Deijl, W. (2025). 'Sentientism and the Welfare Level View.' *The Philosophical Quarterly*, 76 (2): 793–803.
- Whiting, J. (1986). 'Friends and Future Selves.' *The Philosophical Review*, 95 (4): 547–580.
- Windt, J. M. (2020). 'Consciousness in Sleep: How Findings from Sleep and Dream Research Challenge Our Understanding of Sleep, Waking, and Consciousness.' *Philosophy Compass*, 15 (9): 1–16.
- Wolf, S. (2010). *Meaning in Life and Why It Matters*. Princeton: Princeton University Press.