

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Doctoral Dissertations, 2011-2026

Philosophy

2023

Explaining the Distinctiveness of Persons

Register, Christopher

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp01pn89d986f>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

Explaining the Distinctiveness of Persons

CHRISTOPHER C. REGISTER

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF
PHILOSOPHY

ADVISERS: MARK JOHNSTON AND GRACE HELTON

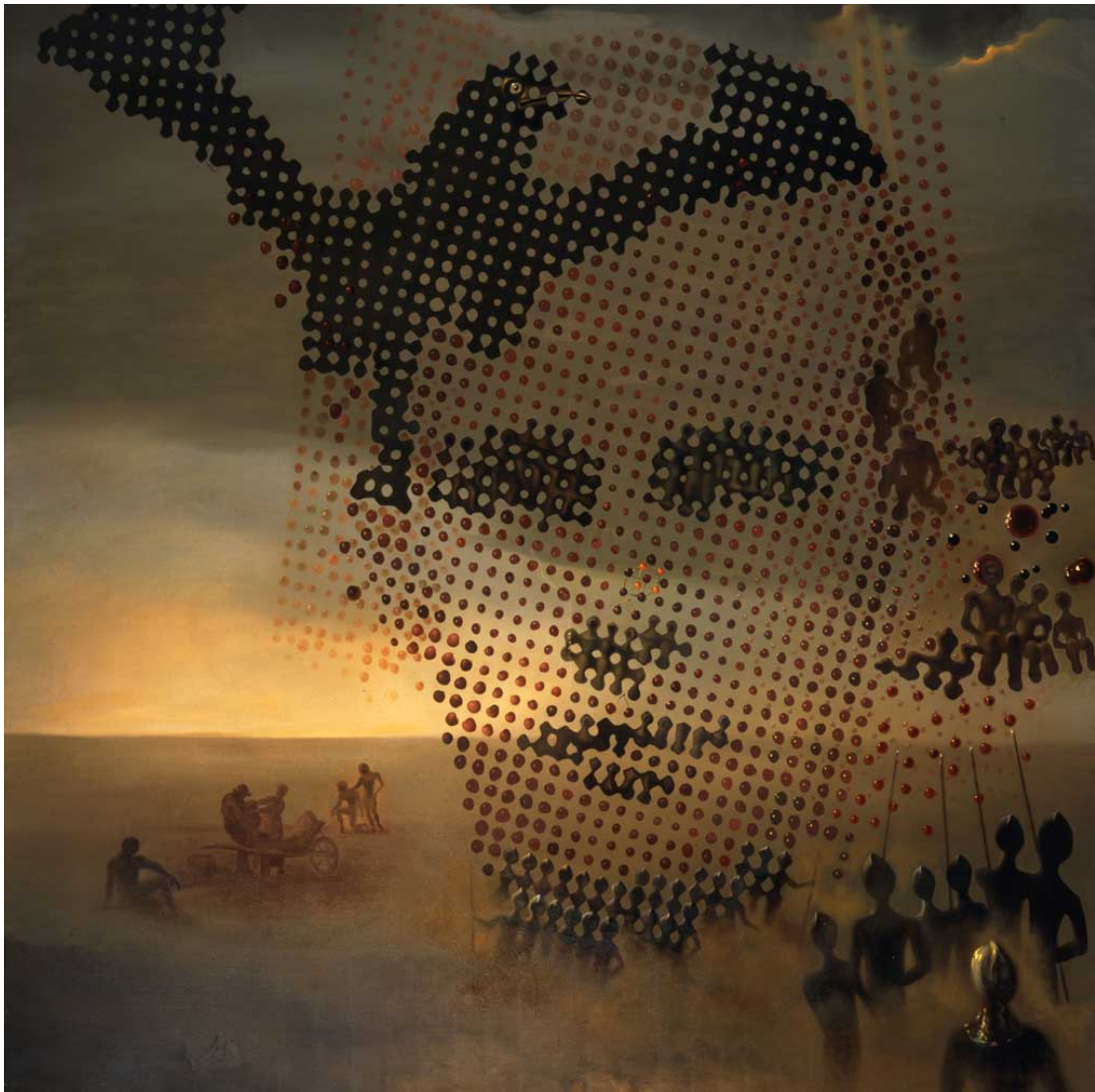
MAY 2023

© COPYRIGHT BY CHRISTOPHER C. REGISTER, 2023. ALL RIGHTS RESERVED.

THESIS ABSTRACT

I propose a new explanation of the moral significance of persons and other morally considerable beings. The explanation appeals to a being's capacity for reflexive, valenced mental states that ground the content and weight of that being's interests. In more sophisticated forms, such mental states acquire first-personal representational content, rising to the level of what I call *self-concern*. Self-concern, I argue, is essential for explaining the distinctiveness of persons.

In this dissertation, I motivate the self-concern explanation by showing how it solves a recent puzzle in the metaphysics of persons. I then defend closely connected theories of personal identity from a challenge posed by the cognitive psychology of the self-concept. After developing and defending this theory of why we matter, I show how it enables a sentientist explanation of the moral significance of permanently unconscious humans. Finally, I draw from the theory to develop and begin solving a new problem in the ethics of artificial intelligence.



Salvador Dalí, *Portrait of My Dead Brother*, 1963, oil on canvas.

Collection of The Dalí Museum, St. Petersburg, FL (USA) 2023
Worldwide: ©Salvador Dalí, Fundació Gala-Salvador Dalí, (Artists Rights Society), 2023
In the USA: © Salvador Dalí Museum, Inc., St. Petersburg, FL, 2023
Photo © David Deranian, 2021

Contents

THESIS ABSTRACT	iii
o INTRODUCTION	i
1 WHY DO WE MATTER?	8
1.1 In Need of an Explanation	9
1.2 The Problem with the Common Explanation	12
1.3 Sortalish Self-Concern	18
1.4 The Diachronic Structure of Self-Concern	21
1.5 Explaining Why We Matter	28
2 THE SELF AND THE CONFLICTED SELF-CONCEPT	34
2.1 Concept and Conflict	35
2.2 Thinking about the Self	38
2.3 The Problem of Conflicting Attitudes	41
2.4 Pluralism	46
2.5 The Self-Concept is not (merely) Dichotomous	49
2.6 The Integrated Self	55
3 SENTIENTISM AND PERMANENT UNCONSCIOUSNESS	61
3.1 Beyond the Horizon	62
3.2 Of Squirrels and Infants	65
3.3 Diachronic Dispositional Sentientism	70
3.4 Interlude on Theoretical Motivation and Coherence	78
3.5 Infants, Persons, and Permanently Unconscious Humans	81
3.6 Moral Status Revisited	85
4 HOW WILL ARTIFICIAL INTELLIGENCE MATTER?	87
4.1 Artificial Interests	88

4.2	The Problem of Artificial Individuation	90
4.3	Why Current Theories Don't Help (Much)	95
4.4	The Sentence-First Approach	103
4.5	Applying the Approach to Cases	110
4.6	Limitations	116
5	CONCLUSION	117
APPENDIX A NOTES ON A METAPHYSICS OF PERSONS		120
A.1	Overview	121
A.2	Toward An Essentially Self-Concerned Being	122
A.3	Dissolving a Problem of Determination	137
A.4	Person Essentialism and Moral Status Essentialism	140
REFERENCES		154

FOR KOINONIA, FOR HALEY.

Acknowledgments

The process of writing a dissertation is often strenuous, and in my own case that was true for largely unpredictable reasons. Changes in the global political climate—not to mention the more literal climate—has taken a toll on me. That started even before the COVID-19 pandemic. Yet, while it can be hard to do good philosophy at a time like this, I am immensely grateful to have had the opportunity to pursue something rich, puzzling, and foundational. For that opportunity, I thank Princeton University, the Department of Philosophy, and my support network—for making this path, and for keeping me on it.

I doubt I would have gotten here without Laurie Paul, whose guidance during and since my time at the University of North Carolina at Chapel Hill has been utterly formative. I was lucky, also, to be at UNC when Laurie published her *Transformative Experience*. While studying the practical hurdles of life's transformations did not help me predict what graduate school would be like, it has helped me make sense of so much of the journey.

A proleptic reason for my wanting to come to Princeton was Mark Johnston's work on the philosophy of mind and personal identity. I have been lucky to learn from Mark over the last several years, to confront his ideas, and to have wrestled with the disquieting problems of personal identity that he has revealed. Much of this thesis is inspired by his early work and forms a response to his later work. In that way, my proleptic reason has been borne out in an uncharacteristically predicted way.

One of the changes that transpired during my time at Princeton was the arrival of Grace Helton, whose presence in this department has been an unparalleled positive for me. Grace's consistent and insightful feedback, her encouragement, and her tactful guidance has been a lighthouse for me. I cannot thank her enough for her support. To be sure, she is an intellectual and academic role model for me.

Several other colleagues in this department or others, faculty and graduate students alike, have been vital sources of inspiration and support over the last few years. So much could be said, but I will simply thank Gideon Rosen, Sarah McGrath, Adam Elga, Liz Harman, Boris Kment, Tom Kelly, Ben Morison, Hans Halvorson, Tania Lombrozo, Ram Neta, Kyle Landrum, Eliya Cohen, Snow Zhang, Joy Shim, Brett Karlan, Claudia Yau, Alisabeth Ayers, Aarthi Vaidyanathan, Thomas Lambert, Joseph Moore, Daniel Krantz Binder, Clau-

dia Dumitru, Alex Meehan, Patrick Miller, Gabe Shapiro, Adele Watkins, Katie Rech, Calvin Baker, and many others. It's been quite the time.

Everyone in our department knows that our beloved administrative staff keep the ship on course. Jo Kelly, Nancy Groll, Anna Faiola, Joann Zuzcek, Jason Robinson, Daniel Rusnak, Anna Colasante, Keisha Craig, and Brittany Masterson: Thank you for everything.

A special heartfelt thanks goes to my close friend Kyle Landrum. His presence in my life has been a gift, not in the least because of our shared philosophical interests, sense of humor, musical tastes, and time spent lifting heavy objects. He has made the pains and gains of graduate school more bearable, and more meaningful too.

Of course, my support network extends beyond people I have met in academic philosophy. First and foremost, my love and appreciation goes out to my spouse and partner, Marleina Ubel, and to our daughter Haley. Without them, I think I would have fallen apart a long time ago. Taking walks, playing videogames, or making and sharing food, with and for my babies, kept me going at the worst of times. I've learned that it is good to have one's center of gravity outside of oneself. (I should say, it's also good to have a therapist.)

Finally, I'd like to thank my parents, Tom Register and Carol Shively, and my sister, Hannah Register. I can't overstate how much freedom, privilege, love, and support they have provided me over the last (nearly) three decades. Being able to share my woes as an academic researcher with all three of them has been a bonus and a comfort.



Introduction

A SQUIRREL, A BABY, AND A SENTIENT ANDROID WALK INTO A BAR. They all flinch from the painful surprise, and the baby starts to cry. It was a metal bar, after all, and the world is worse for it. In fact, because you left the bar there, you have inflicted harm on them—a morally bad thing to do.

Why is this event morally significant? Well, it's worse for each of these creatures to be in pain than not to be in pain. That is, they have an interest in not being in pain, all else equal. And these beings are morally considerable—they have “moral status”. Even though you did not know them and don't owe them any special treatment, they still deserve to have their interests taken into consideration by you and everyone else. When, out of culpable neglect, you harm them by frustrating their interest in not being in pain, you do something wrong. And, following a blameworthy harm of that sort, we hold that some kind of repair is called for. In this case, an apology may suffice.

That explanation may be good as far as it goes, but it doesn't go as far as we'd like. When we try to spell it out in much more detail, it begins to unravel. The problem indicates a flaw in the foundation of our theory of moral status. In this thesis, I diagnose the problem, develop a solution to smooth things over, and then draw out a few surprising conclusions to follow.

A theory of moral beings should tell us, at the very least, which entities have moral status. Ideally, it would give us a principle that explains why those beings have morally considerable interests. That is, it should tell us, quite generally, what the *grounds* of moral status are. There are multiple plausible theories of the grounds of moral status. Some theories hold that a basic mental capacity, such as having the capacity for conscious experience (consciousness) or the capacity for positively or negatively ‘valenced’ conscious experience (sentience), is necessary and sufficient for having moral status. Other theories hold that something more sophisticated, such as being rational or having the capacity for self-conscious thought, is necessary for having moral status, or at least for having the ‘full’ moral status that supports rights. Still other theories appeal to a creature's species, community, origin, potential, or soul—but

I set all of this last group of theories aside, as I do not believe those factors are relevant to whether one's interests are morally considerable.

An initially compelling theory is that sentience is necessary and sufficient for, and grounds, moral status. We can call this view sentientism. Following many theorists, I also hold that 'personhood', the capacity for rational and self-conscious reflection, contributes additional significance to the interests of sentient beings. For the most part, I will focus my attention on revising this view in light of its foundational issues and teasing out the ramifications of that revision.

The problem for all of these theories is that when we try to apply the principles of the theory in a consistent fashion, we end up committed to an unworkable mess of moral demands. The initial goal of the theory is to try to find some guidance for respecting all creatures that matter, but the result is a picture in which the world is rife with moral harm that is systematically unnoticed.

The problem is derived roughly as follows. First, we establish some principle about which beings have moral status. Let's take the sentientist principle. Then, we apply the principle consistently and in full generality: for any being that is sentient, that being's interests are taken to be morally considerable. Finally, we look at the condition of all and only those creatures to determine what moral demands flow from their existence.

Ordinary applications of the theory seem to go well. Is a baby sentient? Almost certainly. So its interests matter morally. The squirrel? Yes. The sentient android? Clearly. Then we might sort out what their interests are. Presumably they do not want to be in pain. Presumably it is wrong to impose costs or harms on them, except in the case of punishment or when exacting reparations from them. Presumably we should not kill them without good

reason. And so on.

The problem arises when we take off the ontological training wheels. For, it's not just babies, squirrels, and robots in our midst. There are also many other objects, such as the cerebrum of the baby, the mereological composite that is the top-half of the squirrel (still attached to the bottom half, but considered as a distinct object), the first core of the central processor of the android, and so on. These are all objects that exist, and of them we must ask: do they satisfy the conditions for having moral status? And in each such case we must respond: yes. That is where the theoretical trouble begins.

If it is wrong to end the life of a sentient thing, then we should not ever replace the first core of the android's CPU. For, that core is a sentient entity, and to replace it would end its activity once and for all. Perhaps more pressingly, it would be wrong to cut a hair on the squirrel's head because that would destroy the top-half composite—thereby ending the life of a sentient being. And more pressingly than that, the growth of the infant seems to involve the death of a great many morally considerable objects that are outside the ontological lanes of ordinary thought. Is getting older *such* a tragedy?

The source of the problem is that we have built our moral theory—which aims at being fully general—upon a foundation of less-than-fully-general ontological presuppositions. These ontological presuppositions, which are assumptions about which objects exist, may well be false. There are many good philosophical reasons to think that the true ontology (the true catalog of existing things) must be quite different than what we ordinarily believe. Either there aren't any complex objects, or there are many more complex objects than we thought. Every other view seems to rely on some badly motivated, arbitrary, anthropocentric, or mysterious theory of what things there are. Recent work in philosophy shows that

such ‘conservative’ ontologies are not really tenable. Yet, once we take off the training wheels of folk-ontology, we see that our moral principles, as stated, become basically unworkable.

The overarching goal of this dissertation is to find a cogent replacement or revision for these unworkable principles, and to draw out some of the consequences of the revised view. I ultimately defend the view that what is distinctive about persons is that we are self-concerned objects, a thesis that is motivated by the moral importance of sentience and self-reflective thought. In chapter 1, I address the source of the problem and motivate a revised theory of the grounds of moral status. This chapter explains how there is typically just one morally significant being overlapping with any person, despite the fact that there are many sentient, thinking entities that partly or completely overlap with me and you and everyone else. I discuss the problem in the context of recent work of Mark Johnston, who has developed the general ‘fellow-traveler worry’ and the more specific ‘personite problem’, according to which there are many morally considerable beings that coincide with any person for part of the person’s life. The basic solution is to appeal to morally enfranchising mental states that are selectively directed at the person and which are not directed at those other unusual overlapping things (like my top-half, or the left hemisphere of my brain). The relevant mental states are those of *self-concern*: an array of dispositions to think and feel certain things about oneself, i.e., a way of caring about oneself.

In chapter 2, I discuss and defend a theory that takes something like self-concern to be a defining condition of personal identity. I discuss the views under the heading of ‘attitude-dependent’ theories of personal identity because the issues I grapple with apply to many other theories in the vicinity. The goal there is to defend such views from the worry that our self-directed attitudes may not really be unified or coherent, in which case a derived theory of

personal identity would be fragmented or incoherent. Against this possibility, I argue that our attitudes are governed by a self-concept that is disposed toward unity and coherence.

In chapter 3, I begin from a sentientist conception of the significance of moral beings to show how that view ultimately points toward the self-concern explanation of chapter 1. I then show how these resources can deliver the verdict that permanently unconscious humans, such as those in permanent comas or who are brain-dead, nevertheless are morally significant in much the same way that typical adult human persons are. I also suggest that the sketch I provide there supports a full fledged theory of the moral significance of moral beings at all stages of life. The first chapter provided a theory of which creatures have moral status; the third chapter works toward a far more ambitious theory of the content and weight of morally considerable interests.

The resources developed in the first three chapters are put to further use in chapter 4, where I apply them to the question of the sentient android—or really, to the question of artificial sentience however it might be embodied. I begin that chapter by developing a worry about the possibility of artificial sentience: how can we know what the interests of artificial beings will be if we do not know how to individuate such beings? Individuation conditions determine the extent of one's body, what it takes to survive, and how many moral creatures there are. If we do not know the individuation conditions that tell us how to separate artificial sentient processes into distinct moral units, then we face serious moral risks of inflicting bodily harm or death on some unknown number of beings. These are possibly catastrophic levels of moral ignorance.

While I do not completely solve the problem in any exhaustive fashion, I do motivate an approach (that I dub the sentience-first approach), which holds promise in solving the

problem of individuating artificial beings. The basic strategy is to look to sentience-involving dispositions to identify the interests and individuation conditions of artificial moral units. The resulting view is similar in spirit to the approach taken in the human case and employs many resources that have been explored earlier in the dissertation.

Finally, in the appendix, I state a theory of the metaphysics of persons—what is basically a version of the attitude-dependent approach discussed in chapter 2. The theory describes some of the more abstruse commitments of the view and tries to show how the various metaphysical commitments are consistent. I do not provide much argument there; rather, the goal is simply to state what has elsewhere been implicit.

Though there are many connections across chapters, each chapter is written as a self-contained unit so that a reader should be able to understand any chapter on its own. That each chapter is self-contained means that there is also some repetition throughout.

This dissertation is part of a larger project that I will continue to pursue. Before my first-ever philosophy class, I became especially interested in the technological prospects of sharing neural or other psychological connections across human minds. When I started at college, I sought the tools of philosophy to think through the questions raised by this possibility. What is death in a world where others will remember from the inside what you saw and felt? What is anger, or shame, when it is not determinately true that the target of your feelings is you or not-you? What kind of ethics or politics makes sense—and makes good—of mind-edness that is shaped like that? In future work, I'd like to extend the framework developed in this dissertation to theorizing about the future of humans and other moral beings.

There are souls, he thought, whose umbilicus has never been cut. They never got weaned from the universe. They do not understand death as an enemy; they look forward to rotting and turning into humus.

Ursula Le Guin, *The Dispossessed*, 1974

1

Why Do We Matter?

We commonly explain the distinctive prudential and moral status of persons in terms of our mental capacities. I draw from recent work to argue that the common explanation is incomplete. I then develop a new explanation: we are ethically important because we are the object of a pattern of self-concern. I argue that the view solves moral problems posed by permissive ontologies, such as the recent personite problem.

1.1 IN NEED OF AN EXPLANATION

PERSONS ESPECIALLY MATTER, both prudentially and morally.¹ Persons are the referents of our pronouns, the relata of our interpersonal relationships, and the default focal points of our practical reasoning. We assume that our practical and evaluative focus on *persons*—rather than brains, bodies, or hedonic pleasure—is justified and not globally criticizable. It would be shocking, then, if it turned out that we were making a radical error in treating persons as distinctively important.

Not only do we assume that our person-centric practices are beyond global reproach, but we think there is a *demand* for us to treat persons well. Persons deserve to be treated a certain way, and failing to treat persons that way is a moral error. We hold that persons are a source of reasons such that, if we fail to appropriately respond to those reasons, then we are doing something wrong.

These are heavy issues, and their heft demands an explanation. Why do persons matter? What is the source of the special moral and prudential significance of persons, and why are other objects (like brains, mannequins, or parts of persons) not similarly significant? What's so special about *us*?

A common explanation cites our mental capacities to explain why we matter.² Persons are more important than mannequins because persons reason, feel, or care while mannequins do not. A person matters more than an ant because we are more rational, more sentient, and

¹I will tend to use the term 'person' loosely. I allow but do not presume that persons have (positive) moral status contingently. My account is intended to explain the prudential and moral significance of non-human and human persons alike.

²See, e.g., Tooley 1972, Singer 1993, Harman 1999, McMahan 2002, Jaworska 2007.

more caring than ants are. However, this explanation only goes so far. It doesn't explain why brains, the top halves of humans, and other unusual parts of persons don't matter in the way persons do. After all, brains, the top half of humans, and many other parts of persons are mentally sophisticated in the way persons are. These odd objects share our brains, and so they also seemingly share our mental capacities to think, feel, and care. Yet they do not matter like we do, and so our explanation falls short.

Trouble for the common explanation arises if these unusual objects exist and share our mental capacities. These claims are widely held, but they are not universally endorsed. Some have argued that at least some of the odd objects are not conscious because they are not maximal (Sider 2003). However, this view requires that consciousness is extrinsic, which is highly tendentious. And, while the permissive ontologies that recognize the existence of these unusual objects have been denied, there are powerful arguments in their favor.³ So, for the purposes of this paper I will assume that mental capacities such as consciousness are intrinsic to brains and human bodies, and that some permissive ontology is true. It is worth stressing that the problem arises even for modest permissive ontologies. If brains exist, can think, and are distinct from human beings, then the common explanation is not enough.

The failure of the common explanation can be drawn out in dramatic fashion by examining the recent *personite problem* developed by Mark Johnston—a specific instance of the ‘fellow-traveler worry’ mentioned in the introduction.⁴ The personite problem shows that

³For proponents of permissive ontologies, see, among many others, Lewis 1986, Yablo 1987, Heller 1990, Hawley 2001, Sider 2001, Bennett 2004, Thomasson 2007, Leslie 2011, and Fairchild and Hawthorne 2018. See Korman 2020 for an extensive review of the literature on restrictive versus permissive ontologies.

⁴See Olson 2010 and Johnston (2010, 2016, 2017, 2021). See also Kovacs (2016, 2020) for discussion of related problems that Kovacs labels *the problem of overlappers*. Kovacs argues for a view of personal identity that explains how we can know we are not any of the non-person overlappers. However, that view does not explain what is ethically special about us.

four-dimensionalist views of personal identity like that of Lewis 1976 are committed to the existence of personites—proper temporal parts of persons. As with other odd objects, the common explanation cannot explain why persons matter in a way that personites do not. The explanatory gap reveals holes in the moral intelligibility of our way of life.

One may be inclined to solve these problems by appealing to a theory of personal identity.⁵ However, the common explanation is incomplete no matter what the true theory of personal identity is. Supposing that human persons are essentially human animals, we must still explain why human animals are ethically superlative (if they are). So, we should first inquire into the conditions that are intelligibly of distinctive ethical importance. Only after should we draw out implications for a theory of personal identity.

This approach—seeking a non-arbitrary foundation of ethical privilege—may also address whether our person-centric practices err in a milder way. Some have recently suggested that some proper temporal parts of persons may independently deserve to be privileged objects of normative discourse (Das and Paul 2020, Dietz 2020). If that's right, then we are systematically failing to appreciate reasons in our midst. What's more, the plausibility of these suggestions shows that we do not really understand whether or why persons are ethically special. While I do not answer the question of whether temporal parts matter independently in some sense, the theory I develop shows why they do not matter like us.

In this chapter, I develop an explanation of the special ethical significance of persons that completes the common explanation. The key innovation is isolating a special mental power that is uniquely possessed by persons and is not possessed by brains, our top-halves, personites, or any of the other odd objects that partially or fully coincide with us. The pro-

⁵Kaiserman 2019 argues that the stage-view avoids the personite problem; however, I think his solution requires resources developed here. See footnote 18.

positional I develop is that an entity is ethically significant in the way we are when it is *the object of a pattern of self-concern*. As developed here, self-concern is a radical overhaul of a classic idea, built for the modern age of permissive ontologies.

In 1.2, I show how the common explanation fails to explain the moral status of persons and then diagnose the problem. In 1.3, I characterize my proposed alternative. In 1.4, I develop a metaphysics of diachronic self-concern stated in terms of counterparts. In 1.5, I show how the new view recovers the distinctive ethical significance of persons, in particular by showing how it solves the personite problem.

1.2 THE PROBLEM WITH THE COMMON EXPLANATION

The special ethical significance of persons has both prudential and moral aspects, but for simplicity I will focus on finding an explanation for why persons have moral status, i.e., why our interests are morally considerable.

Our interests place moral demands on the actions of others. I have an interest in not being hit by a reckless driver. And, if I am hit with a car by a reckless driver, then they owe me compensation. Morally considerable interests come into play twice over in this case, and each way exemplifies the temporal structure of moral status. First, I was harmed in the past, and I am due compensation because of that past offense. Second, while a person is not typically obligated (special relations aside) to provide benefits to someone else, in this case the reckless driver forfeits that sanction. They are on the hook because they harmed me.

How does the common explanation make sense of the fact that I should be compensated? Recall that the common explanation appeals to our mental capacities, such as our ability to reason, feel, or care. Different accounts differ on the details, but we can sketch a

representative answer here. Perhaps I am owed compensation because I suffered as a result of the harm or because I was deprived of something. Or perhaps I am owed compensation because my life was put at risk. These explanations inevitably appeal to the *stake* I have in the events in my life.⁶ My stake in my life is explained, in turn, by appeal to the fact that I am conscious, aware of myself over time, and/or care about my life. Even this brief sketch goes some way toward rendering my deserts intelligible—at least when considered in contrast to the claim that I have no such deserts.

The common explanation involves identity facts: the offended is the same entity as the would-be compensated; the offender is the same entity as the would-be compensator. So, the explanation of why I am owed compensation appeals to two types of facts: my mental sophistication (in virtue of which I can have a stake in things), and my identity over time (in virtue of which I have a stake in particular past or future events). Similarly, the explanation of why the offender owes me compensation appeals to their own mental sophistication (in virtue of which they can be responsible) and their identity over time (in virtue of which they are on the hook for particular past events).

Crucially, according to the common explanation, the fact that X has moral status is wholly explained by X having its intrinsic mental properties—and *not* by X's identity conditions, persistence conditions, or temporal extent. Note that many theorists dispute the notion of 'persistence conditions' and instead appeal to temporal or modal counterparts. To remain neutral on this issue, I'll simply talk of 'persistence properties', which includes the temporal extent of the being, facts about its counterparts, and its persistence conditions (if there are such things). So, X's identity and persistence properties do not figure in that ex-

⁶I have a stake in my life because I stand to "get something out of continuing to exist" (Kamm 2005).

planation. Identity and persistence properties merely delimit the temporal properties of X's interests: they help determine what X's interests are, but they are not part of the explanation for why X's interests are morally significant. As a result, if the common explanation is complete, whether X has moral status is independent from X's identity and persistence properties.

The problem is revealed when we think systematically about moral demands in the context of permissive ontologies. Permissive ontologies characteristically recognize the existence of distinct overlapping objects. Distinct objects have distinct identity conditions and often have different persistence properties. For example, in the classic case of the statue-shaped lump of clay, a permissive ontology typically recognizes the existence of both the statue and the lump, and these objects have different persistence properties. So, if we squash that bit of clay, we destroy the statue but only deform the lump. The lump survives being squashed, while the statue does not. In the same way that the lump overlaps with the statue, there are objects with distinct persistence properties that overlap with persons. There are brains, top-halves of humans, personites, and more. Brains and human top-halves partially overlap with persons at a time. Personites *fully* coincide with persons at a time, though they have different temporal extents. All of these overlapping objects pose problems for the common explanation, but the problem posed by the existence of personites is the most dramatic.

On plausible permissive ontologies, there are many personites.⁷ Personites are exactly like the persons with which they coincide, except that they have radically different persistence properties. They come into and go out of existence frequently. The problem is that, if it's wrong to force someone with moral status to compensate for harms that they did not inflict,

⁷See Eklund 2020 for an additional defense of the existence of personites.

then we face a moral conflict in forcing anyone to pay out for past harms. That's because, for any person who harmed someone, there are many personites coinciding with that person but where the personites did not inflict the harm. These "latter day" personites did not inflict harm because of their temporal extent: they only came into existence after the harmful action. Then, compensation for that harm is both morally required and morally prohibited. Or, at the very least, the interests of the person and many of the personites cannot be jointly satisfied. As Johnston 2017 shows, the level of conflict is massive because of the number of personites.⁸ Moreover, similar conflicts arise when delaying gratification, enforcing promises, taking out loans, and so on, such that virtually any action is riddled with conflict. This is the personite problem.⁹

From a flatfooted perspective, it's preposterous that a personite could have morally considerable interests that conflict with those of the person. The only interests right here are mine! This thought expresses our default presumption of ethical singularity, articulated by Johnston (2017, p. 12) as the principle that:

"The only being with a moral status to be found within a person's spatiotemporal envelope is that person."

⁸Different ontologies disagree on the number of personites. A popular ontology is *liberal perdurantism*, which holds that each distinct spatiotemporal region is exactly occupied by a distinct object. On that view, if time is continuous, there are uncountably infinitely many personites coinciding with any (temporally extended) person at any time.

⁹The personite problem has other aspects in addition to rampant conflicting moral demands. One issue is that of *infinitary paralysis*, which shows that consequentialism does not avoid the problem (Johnston 2017, p. 635-641). If time is continuous, then it turns out that there are uncountably infinitely many personites. Nearly any action whatsoever can then be shown to be infinitely bad and also infinitely good—producing rampant practical impasses. If time is not continuous, then another issue arises: longer lives contain many more personites. If that's right, then against our ordinary sensibilities, longer lives should be heavily prioritized in our ethical reasoning. Fortunately, if we can show that none of these personites (or other fellow-travelers) have what it takes to matter morally, then the three problems of conflicting demands, infinitary paralysis, and the overwhelming priority of long lives are all avoided.

I take it that I have shown how the common explanation cannot establish ethical singularity, even in a modest permissive ontology. Personites share our minds and bodies, so they share all our physical and mental properties. If you can think, feel, and care at some time, then so can any other object that overlaps with your thinking/feeling/caring parts at that time. The common explanation cannot account for *your* singularity. It is incomplete.

On the other hand, it's independently compelling that moral status is explained *somehow* by special mental capacities. Mannequins don't have moral status because they don't have mental capacities like ours. That's true! The question is, how can we extend that idea to explain why personites and other odd objects also lack moral status?

To explain the moral difference between persons and 'odd' objects, we will have to exploit a property of persons that is not had by the other partially or fully overlapping entities. Following Fairchild 2019, I'll call a property that is necessarily shared by fully overlapping objects *neutral*, and I'll call any property that distinguishes such objects *sortalish*. As an example, HAVING MASS *M* is a neutral property because the lump of clay and the coinciding statue necessarily share their mass. In contrast, BEING A WORK OF ART is a property had only by the statue, so that property is sortalish. (The property depends on the *sort* of thing the statue is.) The common explanation fails because the mental capacities it appeals to are neutral¹⁰, so they are shared by coinciding objects. The key here is to find a morally relevant feature that is appropriately sortalish, i.e., that distinguishes persons from the other objects.

More specifically, I propose that we begin with the classic Lockean powers of personhood and adopt a nearby sortalish strengthening of those powers. Inspired by Locke's

¹⁰In fact, mental capacities are what we might call *super-neutral*: they are necessarily shared by objects that share thinking *parts*. That's why the common explanation fails even for partial overlappers. The sortalish property I appeal to distinguishes persons from both full and partial overlappers.

influential definition of a person¹¹, the Lockean powers are powers of self-conscious reflection. Elsewhere, Locke discusses ‘concernment’ as important to what we are, so it’s natural to include not just self-reflective thought, but also ways of feeling and caring.¹² Putting these together, we have *self-concern*: a way of thinking about and caring for oneself first-personally.

That characterization is not yet suitably sortalist. There is a sense in which personites, brains, and human top-halves have such abilities. They can look at themselves in the mirror and think “There I am” in a seemingly uncriticizable way. On the other hand, persons seem to have a privileged relation to the patterns of self-concern they instantiate. For example, many first-personal beliefs that the non-person overlappers token will be false about them. Some of the parts of myself that I care about are not parts of my brain. This is a clue that we are on the right track: there is something about self-concern that promises to be suitably sortalist.

The approach is to locate, within our attitudes, a rich pattern of self-concern that is highly selective. But we will have to do some work. For example, it’s not enough to appeal to our capacity for first-personal thought. The notion that first-personal thought refers reflexively to the thinker is not sufficiently selective because brains and personites are also thinkers. It’s also not enough to appeal to the stronger semantic claim that ‘I’ refers exclusively to the person that tokens it.¹³ Rather, we must appeal to a sortalist mental capacity that plausibly imbues *only* persons with their distinctive ethical significance.

¹¹Locke defines a person as “a thinking intelligent being, with reason and reflection, that can consider itself as itself, the same thinking thing, at various times and places” (Locke 1975).

¹²See, e.g., 4.2.14 of Locke 1975. See also Strawson 2011, especially chapter 4 for a discussion of Locke’s notion of concernment. On the relevance of caring for moral status, see also Jaworska 2007 and Jaworska and Tannenbaum 2014.

¹³Noonan 2010 defends this semantic claim to solve the thinking animal problem. The main problem with this view is that it’s just false that ‘I’ always refers to the person. Sometimes we use ‘I’ to refer to avatars, short-lived selves, or imaginary characters. See Parfit 2012 for relevant discussion. A second issue is that we can introduce new terms that refer to overlapping objects (Johnston 2017). I discuss the latter point in section 1.5.

1.3 SORTALISH SELF-CONCERN

The relevant notion of self-concern is a special way of caring for an object that is *directed selectively* at that object and *ascribes value* to that object.¹⁴ Like any sort of concern for something, self-concern is a pattern of interrelated attitudes and dispositions.¹⁵ Self-concern is different than other kinds of concern, such as concern a child might have for a stuffed toy, because of the distinctive affect, phenomenology, and functional profiles of the attitudes and dispositions involved.

The proposal is that being self-concerned is a sortalish property of persons that explains our distinctive ethical significance. The sortalish nature of self-concern arises from its being selectively directed at a certain sort of object. The evaluative and reason-giving weight of self-concern arises from the value it ascribes to its object. I will now provide a substantive characterization of four expressions of self-concern to show how the two components (selective directedness and value ascription) work.

First consider a simple manifestation of self-concern: your reaction to stubbing your toe. When you stub your toe, there are mental states that collectively ascribe negative value to that event. These mental states have two kinds of components: felt components that are the source of value of the event; and representational and/or functional components that

¹⁴I use the term ‘object’ loosely. The object of self-concern could be a process, a sum of stages, an enduring subject, a human being, or something else.

¹⁵This notion of self-concern is inspired by similar notions from Johnston 1989 and Martin 1997. See also Korsgaard 1989, Rovane 1997, Schechtman 2001, and Johnston 2010, especially chapter 4, for theories that take similar attitudes to play a role in determining personal identity. I don’t have space to discuss differences between these views and mine, but a few differences are worth noting here. First, the views of Korsgaard and Rovane appeal to a *normative* standard of rationality, which I do not think is required. Second, unlike the views of Martin and Schechtman, I do not think that self-concern requires that the subject can accurately anticipate their mental states at other times. In both cases, these requirements are too demanding to be met by ordinary people in many desired cases.

are the source of the selective directedness of the mental states. The felt components involve phenomenology, affect, and motivational aspects, such as pain or an urge to recoil. The representational and/or functional components involve your representations of your toe *as* a part of you, having intentions to move your toe or protect it, and so on. Importantly, the mental states directed at your stubbed toe are an interconnected part of the broader web of self-concern. The toe-directed mental states contribute to the selective directedness of the web. In particular, your attitudes about your toe are (some of) the attitudes in virtue of which the object of self-concern is determined to be something *with that toe as a part*.

A second example of self-concern is fear of anticipated future experiences, such as if you are nervous about an upcoming public speech. Anticipatory self-concern attitudes, such as excitement or fear, have as their object one's future condition. The attitudes ascribe positive or negative value to one's expected experiences. The ascription of value depends on you taking a certain future person (or entity) to be you. So, if you believed that the future person who will experience that event is not you, you would not be afraid. This way of ascribing value is a determinate way in which you are *invested* in, *concerned* for, and *identifying* with that future person. And the conditions that a future person has to meet in order for you to have such attitudes contribute to the selectivity to your pattern of self-concern. If you care for what your life will be like during that public speech, then the future public-speaking person is included within the purview of your self-concern.

A third expression of self-concern is the anger one might feel about a past event, such as if you are angry about having been insulted at a party. While fear or anxiety are *prospective* attitudes, anger tends to be a *retrospective* attitude. These types of expressions of self-concern help delimit the temporal extent of the interests of a person. Retrospective attitudes

often have a special relation to memory. When you remember having been rudely insulted at a party, the phenomenology of that episodic memory may bring along (by psychological default) a sense of having been *that* person in *that* situation. The episode may also elicit an intimate first-personal ‘reliving’ of the feelings one experienced during the original event. The attitudes and feelings that accompany the memory are directed selectively at, and ascribe value to, that past person and the events it lived.

A fourth and final example of self-concern is the excitement or aversion one might feel about imagined possibilities, such as how one might feel at the imagined possibility of becoming a parent.¹⁶ Whereas conscious awareness of one’s body contributes to the mere-ological purview of one’s self-concern, and whereas prospective and retrospective attitudes contribute to the temporal extent of one’s self-concern, modal attitudes delimit the modal profile of one’s self-concern. Such attitudes make it the case that one is not self-concerned with the especially *modally fragile* entities that coincide with oneself, such as the object that overlaps with you but which cannot transform into a parent.¹⁷ If there is a necessarily child-less entity that overlaps with you, then it is not the object of your self-concern because your self-concern is not selectively directed at it.

These aspects of self-concern each contribute to its selective directedness. Because self-concern is selectively directed, it induces the sortalish property of BEING THE OBJECT OF SELF-CONCERN. If we understand the property of BEING SELF-CONCERNED as requiring

¹⁶This sort of attitude is explored empirically and theoretically in McCoy, Paul, and Ullman 2019.

¹⁷An entity that could not survive becoming a parent may seem especially exotic, metaphysically. However, given that becoming a parent could be psychologically transformative for some people (Paul 2014), and given that it is natural on a neo-Lockean view such as that of Parfit 1971 or Lewis 1976 to define a person’s survival conditions in terms of permitted degrees of psychological change, it’s not so exotic to suppose there could be entities defined in this manner that could not become a parent. Compare also to the modal fellow-travelers of Johnston 2019. My proposal is to say that, if a pattern of self-concern does not include such counterfactual conditions, then those conditions are not the identity conditions of the person.

BEING THE OBJECT OF A PATTERN OF SELF-CONCERN, then being self-concerned is also sortalish. Most of the objects that overlap with a person, including brains and personites, are *not* self-concerned in that way—a claim I will defend further in section 1.5.

If our concerns are organized around the conditions of personal identity, then persons uniquely (among overlapping objects) satisfy that sortalish property. It's plausible that our self-concern is directed at ourselves *qua* persons, but I try to shore up that claim in the next section. In particular, I address complications that arise if self-concern changes over time.

1.4 THE DIACHRONIC STRUCTURE OF SELF-CONCERN

I've described substantive features of self-concern and some of its manifestations. This section is devoted to getting clearer about the structure of self-concern over time.

Self-concern is meant to explain in how a person is a singular source of ethical significance among the overlappers. Ideally, we can demonstrate that self-concern unifies the parts of a person, over time and at a time, into a singular unit of ethical significance.

Note, however, that ethical singularity can be stronger or weaker, depending on how temporally extended that singular unit is supposed to be. A weaker version holds that, within the spatiotemporal region of a person *at each time*, there is only one being with moral status. The strongest version holds that *within the whole lifetime* of the person there is only one being with moral status. I'll call the stronger principle *lifetime ethical singularity* and the weaker principle *stagewise ethical singularity*.

Stagewise singularity can in principle solve the problems posed by personites and other odd objects. If it's true that there's just one ethically important thing overlapping with any person at any particular time, then there is room to argue that there is no threat of inevitably

violating some entity's interests.¹⁸ For example, it's plausible that there couldn't be both demands to impose certain costs *and* demands to not impose those costs on that entity. On the other hand, our way of life seems to better fit lifetime ethical singularity, so I hope to work toward that stronger principle. The first goal of this section is to show how self-concern could ground stagewise ethical singularity. The second (more tentative) goal is to show how to bridge the gap from stagewise to lifetime singularity.

Self-concern is a causally robust psychological property that unites a person into a singular ethically privileged whole. We can state this condition in terms of privileged parthood and counterparthood relations that are generated by an instance of self-concern.¹⁹ For brevity, I focus on the temporal counterpart relation generated by self-concern. Analogous parthood and modal counterpart relations could also be developed.

In what follows, I use the term 'self' to pick out whatever entity, existing at a time, is an object of self-concern at that time. I use the term 'counterpart' to pick out the beings considered at times with which that self is self-concerned. Then, for x and y , if x , at time t_1 , is self-concerned for y as it is at t_2 , then y is a counterpart of x . Because of how I use the term

¹⁸ Along these lines, Kaiserman 2019 argues that the personite problem does not arise for the stage view of personal identity. Kaiserman argues that because persons are instantaneous on the stage view, there won't be any person-like objects that coincide with yet are distinct from the person. However, this argument does not solve the personite problem because it doesn't secure stagewise singularity. There are two ways stagewise singularity could still be false. First, because stages have diachronic interests, even stagewise singularity requires that there is a unique morally privileged temporal counterpart relation that determines those interests. Otherwise, conflict still arises. Kaiserman only helps himself to that assumption, but it could fail for reasons already canvassed in Johnston 2016. Second, there could be instantaneous person-like entities, distinct from persons, that *partially* overlap with persons. To rule out this possibility, we need a morally privileged parthood relation, too. Otherwise, we have not secured stagewise singularity. I argue that self-concern provides the morally privileged counterpart relation. A similar argument shows that self-concern provides a privileged parthood relation too.

¹⁹ I adopt this way of speaking for clarity and convenience. I have reservations about adopting an ontology based in temporal parts, primarily because mere spatiotemporal divisions do not seem to carve nature at ontologically interesting joints. However, as long as we remember that some counterpart relations are privileged because they reflect real joints in the world, then talking this way should be harmless.

‘self’ and because of the sortalish nature of being self-concerned, the counterpart relation is reflexive.

Being self-concerned is selective. If x is self-concerned, there will typically be many beings coinciding with x that are not self-concerned. That’s because they have mereological, temporal, or modal properties that deviate from the purview of self-concern. For example, you aren’t self-concerned with the whole batch of molecules that constitute you right now because you are only self-concerned for a being that could survive clipping a fingernail, as evidenced by the fact that you do not fear dying by clipping your fingernail.

Retrospectively, for some self S , S ’s counterparts extend into the past to the extent that S is self-concerned for those entities considered at those times. That typically includes past selves from whom S inherited memories, but it may extend further. Many of us care about our infant selves, as evidenced by my being interested in seeing what I looked like at that age, or your being upset at learning that you were dropped on your head. Each such self S_1 , S_2 , and so on, for whom S exhibits self-concern, past and future, will be counterparts of S on this model.

With the counterpart framework in mind, I can give a new gloss to my overall thesis. My claim is that only the self-concerned being S has morally considerable interests at that time, and the temporal extent of S ’s self-involving interests are grounded in the temporal extent of the counterpart relation generated by S ’s pattern of self-concern.

We could stop here, and the account would allow for a revisionary picture of the units of ethical significance. On the view described so far, with counterpart relations fixed by the psychological traits that are possessed at each time rather than over time, it could turn out that interests change from one moment in time to the next. Then the interests that are grounded

in self-concern would be *time-relative*. Interestingly, there is an account of time-relative interests, along with arguments that it delivers compelling verdicts in a range of cases.²⁰ On the other hand, my secondary goal is to show that the approach can support the ordinary person-based notion of interests. That notion presumes lifetime ethical singularity.

To recover lifetime singularity, we need to find a compelling sense in which self-concern encompasses an entire lifetime. In that case, the time-relativity of interests drops out because the self has the same interests over time. If that result can be established, then the privileged unit of ethics is the whole person.

On such a view, the counterpart relation must be an equivalence relation. If self-concern induces an equivalence relation, then for any two selves in a person's life, each is self-concerned for the other. An equivalence relation is reflexive, symmetric, and transitive. Reflexivity is automatic because for *S* to be self-concerned requires that *S* is self-concerned for *S*.²¹ Symmetry and transitivity require more substantive arguments.

I have not yet specified whether an entity's self-concern is wholly grounded in its intrinsic, current psychological traits, or whether its self-concern is also grounded in facts about other times or other objects. Let's call the former notion *t-intrinsic self-concern*. This sort of self-concern includes all of the self-directed attitudes or dispositions of a being that it determinately possesses at *t*.

It's possible that *t*-intrinsic self-concern fails to be symmetric or transitive. For example, I (considered at t_2) am t_2 -intrinsically self-concerned for my infant self who existed at t_1 , but is it true that my infant self was t_1 -intrinsically self-concerned for me (considered at t_2)?

²⁰McMahan 2002. Some people like the view (Degrazia 2007), but others have criticized it (Broome 2004, Liao 2007, Greaves 2019). See McMahan 2019 for replies to critics.

²¹I grant that there is a sense in which my brain is self-concerned with me. On that way of talking, we would say that only the objects (not the subjects) of self-concern matter.

No—for my infant self did not at that earlier time have any attitudes about me, considered as I am now. Further, supposing that my infant self was t_1 -intrinsically self-concerned with its immediate continued existence and that its t-intrinsic self-concern will become less myopic over time, then transitivity fails too. That's because there will be a chain of selves where S_1 is t-intrinsically self-concerned for S_2 , S_2 for S_3 , ..., and S_n for S_{n+1} , but where S_1 is not t-intrinsically self-concerned for S_{n+1} .

In principle, we could stipulate that self-concern proper is the transitive and symmetric closure of t-intrinsic self-concern. However, I believe this thesis gets the moral facts wrong. To see this, consider a case of a very long-lived person (like that of Methuselah as discussed in Lewis 1976). Consider a very long-lived human being who completely changes their psychological traits many times over, but where there is a chain made up of links of myopic t-intrinsic self-concern from the beginning to the end of this life. To assume just on this basis that we have a lifetime singular being of ethical significance is too quick. Rather, some further substantive condition must be met.

Because it is psychologically possible that t-intrinsic self-concern is non-transitive and non-symmetric, then if we are to secure any substantive sense in which self-concern is guaranteed to be symmetric and transitive, we must take a broader view of which facts determine the profile of self-concern. We must allow that facts about S 's self-concern at t_1 are grounded in facts that are not intrinsic to the mental states of S at t_1 . There are two ways to do this. First, we could hold that self-concern is *diachronically determined*. Or, second, we could hold that self-concern is *socially determined*. I defend each in turn, but I won't decide between them. These claims are optional additions to the core approach and are unnecessary for solving the personite problem.

On the diachronically determined version, we secure the symmetry and transitivity of self-concern by holding that it is open and cumulative. Self-concern is open when its present profile is partly settled by future facts. Self-concern is cumulative when its present profile is partly settled by past facts. On this version, it's not the case that my infant self is not self-concerned for me, even if its t -intrinsic self-concern didn't include me. That aspect was open when I was an infant, and it is settled by facts that came later. Because that pattern of self-concern later included me, I am a counterpart of the infant. Along these lines, I propose the following principles to govern openness and cumulation.

Openness: if S_2 at later time t_2 is t_2 -intrinsically self-concerned for S_1 at earlier time t_1 , then S_1 's self-concern includes S_2 .

Cumulation: if S_1 at earlier time t_1 is t_1 -intrinsically self-concerned for S_2 at later time t_2 , then S_2 's self-concern includes S_1 .

Because I am self-concerned for my infant self, openness ensures that I am included in my infant self's self-concern, recovering transitivity and symmetry in that case. Cumulation ensures that, if a past self is self-concerned for me, then I am concerned for that past self. Cumulation recovers transitivity and symmetry in some cases where a future self does not otherwise include some past self in their t -intrinsic self-concern.

If self-concern is open and cumulative, then it will typically be transitive and symmetric. In those cases, it would be an equivalence relation, and there would be just one self-concerned being within the life of a person. Then, lifetime singularity obtains.

However, there are cases where there is reason to reject openness and/or cumulation, in which case transitivity may then fail. Consider the following case. Suppose I am now

t-intrinsically self-concerned with my 60-year-old self but not t-intrinsically self-concerned with the self as it will be afterward, having developed severe dementia. That is, suppose I don't care what happens after I develop severe dementia; let them do as they wish, for it is no business of mine. But suppose further that my 60-year-old self is t-intrinsically self-concerned for the future self with severe dementia. In that case, there may be good reason to suppose that lifetime ethical singularity fails. Perhaps there really are two morally important beings, with two distinct sets of morally considerable interests, that overlap my 60-year-old self and where only one of them would survive with dementia. It has been noted that the interests of my future selves might conflict with current interests in exactly these cases.²² My account provides possible grounds for and an explanation of this discrepancy. So, while the self-concern account allows that there are realistic cases where stagewise interests at one time diverges from stagewise interests at another time, that is a compelling prediction of the view.

The second way to secure lifetime singularity appeals to social determination. This approach begins with the general observation that each individual self inherits a large swath of their self-concern from a social background of person-centric practical reasoning. We tell our children that they should go to school in order to live an easier life overall, we worry about what it will be like for them to care for us in our old age, our parents tell us stories about funny things we did as toddlers, and so on. Webs of questions, habits, and conventions shape our thoughts and feelings.²³ On the whole, the complex web of our practices can help fill out our

²²McMahan 2002, Walsh 2020. There is a converse worry that someone could disavow a past self and thereby not be morally responsible for that past self's actions, on my account. In principle, something like that is possible. However, in practice, my account does not have this implication because human persons cannot adjust their own psychologies radically enough to become genuinely not self-concerned for their past selves.

²³Cf. Adams 1989, who endorses the thought that "the way we care about persons is and should be affected by the deep embedding of the concept of personal identity in a complex web of social practices", and Wolf 1986 who develops a version of that idea. Johnston 1989 also takes the infant's social circumstances to determine its conditions of persistence.

self-concern, giving determinate content to an individual's otherwise indeterminate mental states. These social facts help direct our judgments, values, beliefs, and desires so that they are organized around a sense of ourselves as persisting objects of self-concern. And so, when we reflect on what we care about, we find ourselves thinking about the object at the center of the web, the person, who is the salient target of our self-concern.

On this version of the view, the pattern of someone's self-concern is determined in part by the social fabric in which they are embedded. Then, infants' self-concern need not be indeterminate, even if their t-intrinsic self-concern is indeterminate.

So, we can see how patterns of self-concern could determine an equivalence class of selves, at least in some circumstances. And we have seen that securing at least a time-indexed notion of univocal t-intrinsic self-concern is fairly straightforward. With these options in mind, let's now see how the proposal explains our distinctive, *singular* ethical status.

1.5 EXPLAINING WHY WE MATTER

To see how the condition of sortalish self-concern explains why we especially matter, let's see why odd objects are not objects of self-concern. In particular, I explain with an example why personites aren't objects of self-concern. The explanation shows how coherent moral individuation flows from facts about our psychology.

Consider again the incident where a reckless driver hits me while I am walking. Call the reckless driver Pat. Suppose Pat gets caught a few weeks later and is ordered by the court to pay for my medical bills along with additional monetary compensation. Imagine that the demands require a humbling apology that will be unpleasant for Pat. There are many personites, overlapping with Pat, that came into existence *after* the reckless driving incident and

exist up through and after the payment of compensation. So, many personites have the burden of compensating me imposed on them even though they did not cause the harm. Why does this imposition not violate morally considerable interests?

The short answer, supplied by the proposed account, is that personites do not have morally considerable interests because they are not an object of self-concern. To see why personites are not objects of self-concern, note that there are no mental states located in Pat that are sensitive to those personites. Or, more carefully, there are no mental states that both qualify as expressions of self-concern and are appropriately organized around any of the personites. After all, no first-personal attitudes are sensitive to the coming-into-existence of any of those personites. Moreover, even if you explained what personites were so that Pat could think about some particular personite, that wouldn't bring along with it self-concern for the personite. For example, Pat would not become angry that some personites were made to pay for a harm they did not inflict. So, while it may be possible for Pat to think about *personites*, those thoughts do not come along with the various emotional, affective, and evaluative components that are constitutive of self-concern, and they will not be integrated with other reflexive attitudes. Those thoughts are not functionally integrated in the right way with Pat's other attitudes of self-concern.

At this point, we can address an argument given in Johnston (2017). Johnston notes that it's not defensible to deny an entity moral consideration *merely* because the available language does not have terms that refer to them in a first-personal way. Johnston suggests that we can easily introduce a new term to our language, "wehere", to enable personites to refer to themselves first-personally. "Wehere" is stipulated to refer to all and only the personites that coincide with the person at the time of tokening. Then, personites would have the ability to

self-refer, at least plurally. The ease with which we can remedy the personites' linguistic disadvantage, without changing what they are like in other respects, suggests that their linguistic disadvantage isn't justifiably disenfranchising.

I agree that Johnston shows that *self-reference* is not an ability that makes a moral difference. On the proposed view, the relevant condition is *self-concern*—a rich value-laden analog to self-reference. Does an analogous objection show that personites are just so close to being the object of self-concern, if only they were freed of the limitations of person-centric ideology? If personites are nearly capable of being self-concerned, then it may not be defensible to deny them moral status on that thin basis.

My response to this objection is twofold. First, I claim that it's not psychologically possible for a human mind to instantiate self-concern for *each* of multiple personites. This rules out the possibility of rampant conflicting moral demands. Second, while it is possible for a human mind to instantiate self-concern for just one personite, there are psychological constraints on how that self-concern can be directed. (For example, it's not within your power to be indifferent about what happens tomorrow, but that is what would be required in order for your pattern of concern to be organized around a personite that ceases to exist tonight.) These psychological constraints preclude the possibility of bizarre or wildly conflicting moral demands flowing from personites. Each of these responses plays on the theme that the structure of interests is intimately connected to the structure of self-concern.

The two claims together (to be defended just below) refute the idea that personites nearly possess morally enfranchising mental capacities. These claims show that personites are not on the verge of caring about themselves. Rather, it would take deep psychological transformations to create personite self-concern—transformations that would destroy one

set of interests and bring forth a new set of interests altogether. So, there is no sense in which something with moral status is being unfairly disregarded.

First, I claim that there are constraints, flowing from limitations in our psychology and from the nature of self-concern, that preclude the possibility of consistent and rich *plural* personite self-concern. Plural personite self-concern is where there is, contained within one mind, a plurality of distinct patterns of self-concern, each of which is organized around a distinct personite. As a matter of human psychology, this is not possible because a human mind cannot co-instantiate each of a variety of patterns of self-concern. The states that constitute self-concern directed at one personite crowds out the states that would count as self-concern for a different personite. For example, you can't both fearfully anticipate dying tonight and also be convinced that you will wake in the morning. Or, a little more tentatively, there are drastic limitations to how many patterns of self-concern one could thoroughly instantiate.

I do grant that, if some psychology could be constructed such that there were rich patterns of self-concern organized around a plurality of distinct things, then that would produce a plurality of coinciding entities each with (perhaps conflicting) morally considerable interests. Arguably, that would be the right moral diagnosis. But that's not the sort of restructuring that can be accomplished merely by introducing new terms or ideology.

The core barrier to plural self-concern is the functional cohesiveness of human psychology, which precludes attitudes that are simultaneously strong, pervasive, and conflicting. So, I cannot contain a 'self' that is deeply concerned about dying at midnight on December 31st, 2023, and also contain a self that is excited for going on vacation in January 2024.

Functional cohesiveness can break down. There is a medical procedure where the corpus collosum, the part of the brain that carries neural activity between the brain hemispheres,

is severed. In such cases, it's plausible that patients have divided consciousness. It's plausible they might also develop divergent concerns: perhaps one hemisphere is self-concerned, in some sense, with a prospective future self that has dementia, whereas the other hemisphere rejects that future self. What do we say about this case?

I think the appropriate diagnosis is that there are difficult, uncomfortable, and conflicting moral demands that are generated by this person's psychology. It is genuinely a case where the psychology of a person begins to look more like the psychology of two people. In this case, it may be appropriate to regard this person as having two sets of interests that sometimes diverge. That is a verdict that is explained by my account.

Now for the second claim. I agree that a psychology could in principle be organized around a *single* personite, such as the personite that cannot survive the onset of dementia. But here again, there are limitations that flow from the nature of our psychology and from the nature of self-concern that prohibit some forms of self-concern. It's plausible that a human's self-concern could be consistently and richly organized around conditions that exclude having severe dementia. Then, that person's self-concern extends up to, but not into, that future condition. (Note that ethical singularity obtains in that case.) In contrast, however, self-concern cannot be organized around the arbitrary cut-off points that demarcate the boundaries of many of the personites. For example, some of the personites go out of existence when psychological connectedness is disrupted just a little too much. (Say, if two selves are connected to the tragic degree of .49 instead of Parfit's preferred degree of exactly .5.) The attitudes that constitute self-concern are not sharply sensitive to that kind of fact, and it's hard to see how those attitudes could be systematically sensitive in that way. After all, many memories of past events will persist anyway, and the continuation of those memories brings along

self-concern. So, while it's plausible that conditions other than the expected conditions of personal identity could delimit the purview of self-concern, those are not bizarre cases inevitably riddled with moral harm.

To close, I will address Johnston's argument from duplication (2016, p. 203-204). The argument has three premises, which are roughly that: (1) for any personite, there is an intrinsic duplicate of that personite in some possible world, where that duplicate is a person; (2) moral status cannot differ between intrinsic duplicates; and (3), all persons have moral status. Therefore, because moral status is intrinsic and the personite is intrinsically just like a possible person, then for any personite, that personite has moral status.

I reject premise (2). The premise is implausible because sortalish properties often differ across duplicates, whether or not they are intrinsic, and moral status quite plausibly requires some such sortalish property—as I've argued in the case of BEING SELF-CONCERNED.

If sortalish properties are intrinsic, then intrinsic properties can differ between duplicates.²⁴ (The statue-shaped hunk of clay is a duplicate of a possible statue, but it has different sortalish properties.) Then, a personite and a duplicate could differ in a morally relevant intrinsic respect. After all, the properties that distinguish the statue from the clay, as well as those that distinguish *you* from the hunk of matter that constitutes you, are all sortalish properties. It is plausible that some such property is necessary for moral status.

If, instead, sortalish properties are not intrinsic—preserving the duplication criterion of intrinsicity—then it's plausible that the grounds of moral status are not intrinsic after all (just like BEING A STATUE). That may be a revisionary conclusion, but it's an inevitable implication of the duplication criterion of intrinsicity when there are coinciding objects.

²⁴If needed, we could define duplication in terms of having the same fundamental properties or parts.

What is special about people, about selves, is that what constitutes their identity through time is partially determined by their own conception of themselves, a conception which may vary, perhaps appropriately does vary, from person to person.

Robert Nozick, 1981

2

The Self and the Conflicted Self-Concept

Evidence from cognitive psychology shows that we think about ourselves in conflicting ways. Some philosophers take this to show that there is a plurality of ‘selves’. In this chapter, I motivate an approach to personal identity that supports this inference from how we think about the self to the ontology of the self. However, I also argue against pluralism about selves. To argue for this conclusion, I examine further evidence from cognitive psychology that under-

mines support for pluralism. I then develop a model of the self-concept as a web of imperfectly integrated attitudes, show how it could produce the observed conflicting judgments, and suggest that it could nevertheless still ground the existence of a singular self.

2.1 CONCEPT AND CONFLICT

NO SINGLE THEORY adequately captures how we talk and think about being the “same person” over time. The difficulty is that sometimes we seem to talk one way, but other times we talk another way. For example, sometimes we treat someone as being the same person even if their mind is gone, such as when we encounter a loved one in a persistent vegetative state. Other times, we might say that someone is a “different person” merely if their personality changes significantly, like in the famous case of Phineas Gage.¹

The two prominent theories of personal identity are divided on these cases. The *neo-Lockean* theory holds that ‘person’ is a psychological notion, and that being the same person over time requires psychological continuity.² In contrast, *animalism* holds that human persons just are human organisms, and that being the same person over time requires biological continuity. The neo-Lockean view has trouble allowing that someone could exist in a persistent vegetative state, but it can capture what we mean when we say that Phineas Gage is not the same person he once was.³ Conversely, animalism fails to capture the importance of

¹In 1848, Phineas Gage suffered a traumatic brain injury that reportedly altered his personality in significant ways. The case is often discussed in both psychology and philosophy. See, e.g., O’Driscoll and Leach 1998, Strohming and Nichols 2015, and Tobia 2015.

²This approach owes to Locke’s characterization of a person as a being capable of self-reflective thought and of personhood as the grounds of “forensic” relations such as responsibility. Cf. Locke 1975.

³Neo-Lockean views are compatible with Gage being the same person before and after transformation. What these views capture is that Gage transformed in ways that are relevant to survival and identity. So, if Gage’s transformation had only been more extreme, the neo-Lockean would say he did not survive it.

Gage's transformation, though it can explain the loved one's presence despite their lack of psychological continuity.

The ambivalence in our attitudes is remarkably robust. It turns up again and again in the philosophical literature on personal identity, and psychologists have reproduced the same or very similar conflicts in experimental settings. The data suggests that there is a robust dichotomy in how we think about people. Sometimes we think about persons as human organisms, yet at other times we treat certain psychological traits as essential.

The robustness of this ambivalence has led philosophers and psychologists alike to endorse pluralism, not only about how we *think* about personal identity, but also about what *exists*.⁴ That is, many theorists conclude that talk of being the same person over time runs together multiple distinct things, including biological continuity, psychological continuity, and perhaps other things too. On the pluralist view, personal identity consists in multiple distinct conditions that usually overlap. There is a biological thing and a psychological thing, and they can come apart. Further, these things play distinct roles: the psychological thing grounds relations of responsibility and compensation, while the biological thing grounds relations of survival and anticipation.

However, pluralism does not sit easy with our ordinary beliefs and practices or with first-personal phenomenology. We take ourselves to be unified selves, living unified lives and experiencing the world from a singular perspective. So, there is reason to be dissatisfied with pluralism about personal identity. Indeed, the apparent need for pluralism is sometimes treated as a problem.⁵

In this spirit, some philosophers take the apparent unity of first-personal phenomenol-

⁴Sider 2001, Shoemaker 2007 and 2016, Tierney et al. 2014, Tierney 2020.

⁵See Schechtman 2014, pp. 80-88

ogy, attitudes, and practices not just as evidence about the unity of personal identity, but in fact as the grounds of personal identity itself. Theorists have variously sought to ground personal identity in terms of agential ownership⁶, person-directed attitudes and dispositions⁷, narrativity⁸, conventions⁹, the unity of a ‘person life’¹⁰, and our ‘I-beliefs’¹¹. A common idea is that the person is essentially the focal point of special attitudes or practical roles. Let’s call these views *attitude-dependent* theories of personal identity.

Because we feel and believe that we are unified, singular entities, attitude-dependent theories may help avoid pluralism. However, it’s not clear how that would work. After all, the evidence shows that our attitudes and the corresponding practices give conflicting verdicts in many cases. So, pluralism is not automatically avoided on these views, and it’s an open question whether attitude-dependent views can instead ground monism about persons.

The direct aim of this chapter is to show how an attitude-dependent approach to personal identity could ground monism despite the apparent pluralism encoded in our attitudes. There are two further indirect aims: to motivate and support both the attitude-dependent approach and the recent ‘empirical turn’ in personal identity. I accomplish these aims by examining numerous recent empirical studies that investigate how we think about persons. I ultimately propose that monism is viable on structured attitude-dependent views, wherein the divergent conditions are interpreted as parts of the person. Our attitudes about persons are structured and functionally integrated by the self-concept. Which conditions count as parts of the self, and how those parts are organized into a whole, is partly determined by the

⁶Korsgaard 1989.

⁷Johnston 1989 and 2010, Ch. 4. See also Zimmerman 2013.

⁸Schechtman 1996.

⁹Braddon-Mitchell and Miller 2004.

¹⁰Schechtman 2014.

¹¹Kovacs 2016, 2020.

self-concept.

The chapter proceeds as follows. In section 2.2, I sketch and motivate the attitude-dependent approach. In section 2.3, I review evidence from both philosophy and cognitive psychology that demonstrates the deep conflicts in our attitudes about persons. A key hypothesis is that the conflicts in our attitudes exhibit a robust dichotomy wherein distinct conditions are taken to play divergent practical roles. In section 2.4, I show how pluralism provides a cogent resolution to these conflicts. In section 2.5, I argue that other results speak against the dichotomy, motivating a revised interpretation of the data. In section 2.6, I develop a model of the self-concept that explains the conflict in our attitudes while preserving both singularity and integration.

2.2 THINKING ABOUT THE SELF

I'll begin by sketching and motivating the attitude-dependent approach to personal identity. I will highlight advantages of the general approach over the two prominent alternatives, neo-Lockeanism and animalism.

Generally, attitude-dependent views hold that what it takes to be the same person over time depends on how the individual or the community thinks about and treats persons. Very roughly, these views hold that the conditions that are taken to be the identity conditions of persons thereby are the identity conditions of persons. With respect to identity over time, such a condition might be stated as follows:

Attitude-Dependent Identity Over Time: Person A, considered at time t_1 , is the same person as entity B, considered at t_2 , only if A is regarded as the same person as B.

So, if you think and act as though psychological continuity secures your continued existence, then a future entity that is psychologically continuous with you is in fact you. If, on the other hand, you regard the continuation of your functioning body as securing your survival despite the loss of your distinctive psychological traits, then you could in fact survive losing those traits as long as your body continues to exist. Depending on how your attitudes are organized, different relations could realize the role of personal identity.

Note that the condition stated above is merely a necessary condition for being the same person over time. Typically, attitude-dependent views impose other constraints too. For example, both Johnston (1989) and Kovacs (2020) hold that there must also be at least a minimal degree of continuity over time and some self-reflective psychological traits at some time or other.¹²

There are several advantages of attitude-dependent theories. Some of these advantages are already present in the literature, while others are new.

One advantage is that the view allows for differences across individuals or communities. For example, it seems plausible that there could be a ‘Star Trek’ community: a community of people who think and act as though they can travel to distant worlds by means of teleportation. In contrast, it’s also plausible that there could be a community of people who think and act as though “teleportation” results in death of the original person and the creation of a new person. The people of this community think and act as though survival requires having one’s original human body, and therefore they believe that they cannot be teleported in the manner of Star Trek.

Some have argued that each community is right on its own terms.¹³ The attitude-

¹²See Johnston 1989, p. 457, and Kovacs 2020, p. 350.

¹³See Johnston (1989) and (2010), especially Ch. 4.

dependent approach gives a unified theory that respects these differences. For an individual in either community, part of what it takes for that individual to continue to exist is for there to be a future entity that meets the conditions that the individual regards as their survival conditions.

Allowing for individual differences in identity over time also enables an ecumenical theory of survival in the context of persistent vegetative states. Some people might think that a persistent vegetative state counts as death, whereas others may regard that condition as survival. Supposing that personal identity makes a moral difference, then we can justify differential moral prescriptions depending on the personal beliefs of the patient. For a patient who does not believe they survive in a vegetative state, there is no strong moral reason to keep their vegetative body on life support. For patients who believe otherwise, there is.¹⁴ The advantage of the attitude-dependent view is that it provides a satisfying justification for patient-relative moral claims.

The attitude-dependent view has advantages at the beginning of life, too. Neo-Lockean views have trouble explaining how you could have been younger than an infant, given that an infant has so few distinctive psychological similarities with you.¹⁵ Animalism, meanwhile, entails that you existed as a collection of barely differentiated cells (because the human organism exists at that stage). It's natural to think that our identity conditions are somewhere in-between. Maybe you came into existence not long before birth, or at birth. We can explain that fact by appeal to some attitudes and/or practices, either yours or those of other members

¹⁴I develop a version of this view in Chapter 3. There, I motivate the view by defending an extension of the widely held view that sentience is necessary and sufficient for moral status.

¹⁵Neo-Lockean views are compatible with the fact that you were an infant only if they do not require any sharing of particular psychological connections over time. However, those versions of the view allow counter-intuitive examples of survival through drastic change. Cf. the discussion of Methuselah in Lewis 1976.

of your community. If you treat that newborn predecessor as you, then in fact it is you. In this way, the attitude-dependent approach gives cogent verdicts that are unavailable on the standard views.

The foregoing examples show advantages of attitude-dependent views involving the survival conditions of persons. Those conditions involve what it takes to continue to exist over time. Theories of personal identity should also specify what constitutes a person at a time. Here too, the attitude-dependent approach has advantages. For example, we can allow that one individual adopts a prosthetic limb as an integral part of their body, whereas for another individual it counts merely as a tool. The relevant difference here plausibly stems from how each person psychologically relates to their prosthetic. One person has an array of attitudes directed at the prosthetic that makes it an especially important part of them, whereas the other person does not. The attitude-dependent view delivers this verdict. Other views about the metaphysics of persons are either silent on or incompatible with this verdict.

We've seen how attitude-dependent views can give compelling verdicts about identity over time or about constitution at a time. We could provide analogous applications in the context of modal properties, i.e., the properties that say what a person could be or how they could have been. For now, I will set those other applications aside. The rest of the chapter focuses on the traditional centerpiece of personal identity: existence over time. Unfortunately, it turns out that we have conflicting attitudes about what secures our survival.

2.3 THE PROBLEM OF CONFLICTING ATTITUDES

I've given reasons to take the attitude-dependent approach seriously. In this section, I draw from empirical evidence to develop a problem for the view. This problem motivates plural-

ism, which I address in the subsequent sections.

The problem faced by attitude-dependent views is that our attitudes are not well-behaved. Empirical studies show that people give inconsistent judgments about how we persist. The evidence also shows that our judgments about different practical relations, like responsibility and anticipation, seem to track different conditions. These results pose a *prima facie* problem for the attitude-dependent approach. In particular, it's not clear how to extract consistent facts about the nature of the self from inconsistent attitudes.

The problem of conflicting attitudes goes back to classic thought experiments from Williams (1970). Williams develops two thought experiments that describe the same situation in two different ways, and he shows that we have conflicting intuitions across these cases. In the first description, we imagine that person A and person B are about to undergo an operation in which the psychological traits of A are put into the B-body and the psychological traits of B are put into the A-body. People tend to have the intuition that the operation is one of body-swapping: after the operation, person A will in fact wake up in the B-body, and vice versa. This intuition reflects a belief that the continuation of psychological traits secures personal persistence.

In the second description, the reader imagines that they ("you") will undergo a procedure that will replace all of their psychological traits with new traits. After the mind-replacement procedure, "you" will then be subjected to intense pain. The typical intuition here is that one expects to experience the pain despite the fact that one's distinctive psychology does not persist. And further, this intuition does not change even if we learn that our own psychology was 'implanted' into another body. This intuition reflects a belief that personal persistence is secured by bodily continuity rather than psychological continuity—directly

contradicting the first intuition.¹⁶

The contradictory intuitions revealed by Williams have since been reproduced in psychological experiments. Blok et al. (2005) had participants read vignettes about “Jim” who undergoes a brain transplant procedure that either does or does not preserve memories. The participants were asked whether the transplant recipient was “still Jim”, and they tended to judge that Jim survives brain transplant procedures only when memories are preserved.¹⁷ These judgments plausibly reflect a belief that retaining some psychological traits is necessary for survival.

In a series of studies, Nichols and Bruno (2010) corroborated and expanded on these results. They first tested the psychological condition, confirming that participants tend to agree that the patient is still Jim after the transplant procedure only if his memories are preserved.¹⁸ Nichols and Bruno then tested the second case of Williams (1970). They presented participants with a vignette about a mind-wipe procedure and asked whether the patient (“Jerry”) would feel pain afterward. In a Yes/No forced choice measure, 72% agreed that “Jerry will feel the pain.” These results confirmed that participants judge that someone will experience pain despite losing all memories.¹⁹ However, in yet another study, participants were asked explicitly: “In order for some person in the future to be you, that person doesn’t need to have any of your memories: Agree/Disagree.” Over 80% of participants disagreed,

¹⁶Ninan (2021) argues that the second intuition actually reflects an endorsement of the simple view of personal identity rather than the bodily continuity view. Given the evidence to follow, that nuance won’t impact my argument.

¹⁷On a 0-9 scale, participants agreed that it was still Jim with a mean response of 6.6 when memories were preserved, versus 2.0 when memories were not preserved.

¹⁸On a 0-9 scale, the mean responses were 5.45 when memories were preserved, versus 3.24 when memories were not preserved.

¹⁹Williams speculated that the 1st vs. 3rd-personal framing of the cases explained the discrepancy. These results refute that hypothesis.

suggesting even more strongly that they endorse the claim that memories are required for personal identity.

The results suggest that there is a dichotomy in our judgments about how persons persist. Sometimes we judge as though psychological continuity is required for persistence, yet other times we judge as though bodily continuity is sufficient for securing persistence.

Using a different experimental paradigm, Tierney et al. (2014) provide further evidence of this dichotomy. Instead of eliciting category judgments about vignettes, Tierney and colleagues examine how different perceived continuity conditions affect *temporal discounting*. Temporal discounting is a ubiquitous time bias where people discount the anticipated value of temporally distant outcomes relative to temporally near outcomes. For example, people tend to judge that receiving \$100 tomorrow is preferable to receiving \$110 in a month (Frederick et al. 2002). The regularity of such judgments suggests that we discount the value of outcomes as a function of their temporal distance. That function is called the *discount rate*.

Bartels and Urminsky (2011) showed that perceived psychological connectedness (i.e., perceived similarity) to future selves affects the discount rate. Using vignettes that manipulate participants' sense of connectedness to a future self, they showed that high perceived self-connectedness reduces the discount rate, whereas low connectedness increased the discount rate. In other words, people exhibited less temporal discounting when they felt more psychologically connected to their future self. This work shows that judgments about the self are involved in practical reasoning about the future, suggesting that our attitudes can be probed with discounting tasks. Additionally, the practical judgments expressed in discounting tasks determine identity on some attitude-dependent views.

Tierney et al. (2014) used this paradigm to test whether the Williams' dichotomy is also reflected in practical judgments. The hypothesis is that if perceived psychological connectedness affects some discounting tasks but not others, then we have further evidence that there are at least two ways of thinking about the self: one that is sensitive to psychological traits, and one that is not. To test the hypothesis, Tierney and colleagues compared how perceived psychological connectedness impacts judgments of *how much punishment you would deserve for cheating in the past* versus its impact on judgments of *how anxious you would be about a future root canal*.

Results showed that some but not all self-involving practical judgments are sensitive to connectedness. In particular, beliefs about psychological connectedness do not impact the anticipated badness of future root canals. The pain is expected to be just as bad no matter how psychologically connected the future self is to the current self.

In contrast, when participants judge the amount of punishment they deserve for cheating, then connectedness does make a difference. In particular, participants judge that a person deserves less punishment when they are less psychologically connected to their past self.²⁰ This dissociation provides more evidence for a dichotomy between psychological and bodily notions of the self.

The studies so far have explored attitudes from two angles: explicit judgments about whether someone persists, and practical judgments that implicate the attitudes about the self. From both angles, responses apparently reveal one and the same dichotomy between a psychological and a bodily notion of the self. Because the same dichotomy appears when probing attitudes from two angles, we have reason to believe that these effects are not merely

²⁰Mott (2018) shows that perceived psychological connectedness affects judgments of legal punishment and moral criticism in similar ways.

superficial artifacts of experimental design.

The robust dichotomy in how we think about the self suggests that we have genuinely conflicting attitudes about what we are. This conflict poses a problem for the attitude-dependent approach. The problem is that, if we are conflicted about what we take ourselves to be, then it's hard to see how those attitudes could determine coherent, univocal conditions of personal identity. The attitude-dependent approach is motivated in part by the hope that these attitudes could adjudicate difficult questions of persistence, such as whether or not an individual can survive amnesia, severe dementia, teleportation, or mind-uploading. Given that we have robust conflicts in our attitudes, it's hard to see how the approach could provide helpful, determinate answers in such cases. That is the problem of conflicting attitudes.

2.4 PLURALISM

Recently, some philosophers have defended pluralist views of the self and personal identity (Shoemaker 2007, Tierney et al. 2014, Shoemaker 2016, Tierney 2020). Some of the evidence taken to support pluralism is the very same evidence discussed in the last section, which shows that our attitudes seem to be organized around a plurality of distinct conditions. Indeed, pluralism about the self provides a solution to the problem of conflicting attitudes. That's because divergent attitudes are only contradictory on the assumption that they are about the same thing. If, instead, we allow that the divergent attitudes are in fact about distinct things, where one thing can exist without the other, then the conflict is resolved. So, there is reason to think that the attitude-dependent approach leads to pluralism. That is *not* my conclusion. However, before assessing pluralism, it will help to get clearer on what the view is and how the data supports it.

Personal identity has often been taken to be about a single practically important type of entity. Closer examination reveals that personal identity apparently consists in a combination of multiple distinct conditions. What's more, these distinct conditions can come apart, leading to what has been called the *problem of multiplicity*.²¹ For example, it's plausible that the continuity of distinctive psychological traits is important for being responsible for past actions and perhaps for deserving compensation for past harms. In contrast, bodily continuity or continuity of the first-personal perspective is important for survival and the rational anticipation of experiences (such as anticipating pain). Given that you can have bodily persistence without the persistence of psychological traits, you can have one of these practical features (rational anticipation) without the other (desert).

Along these lines, Shoemaker (2007, 2016) argues that these distinct practical roles are grounded in a plurality of possibly diverging relations. Similarly, Tierney (2020) develops what's called the Subscript View, according to which there are multiple distinct overlapping entities—selves. The distinct selves are notated by subscripts, e.g., ' S_B ' and ' S_P ', where each distinct self has distinct persistence conditions (bodily, psychological, and perhaps others too). For the purposes of this chapter, it doesn't matter whether we think of pluralism in terms of multiple *relations* or multiple *entities*, so I lump these views together.

Pluralism solves the problem of conflicting attitudes. In fact, the prevalence of conflicting attitudes provides a *prima facie* case for pluralism. Recall the dichotomous regularities in attitudes about selves: some of our attitudes appear to be organized around one type of condition (psychological continuity), whereas other attitudes appear to be organized

²¹See Shoemaker (2007, 2016), Schechtman (2014), Tierney et al. (2014), and Tierney (2020). Multiplicity is a problem because it is counterintuitive and because it erodes the importance of persons and theories of personal identity.

around another type of condition (bodily continuity). The dichotomy provides some reason to think that the concept governing these attitudes, the self-concept, is either plural (i.e., is really multiple distinct concepts) or polysemous.²²

Next observe that our attitudes about one practical relation (deserving punishment) are sensitive to psychological continuity, whereas attitudes about another practical relation (pain anticipations) are sensitive to bodily continuity. This pattern suggests that people think that distinct conditions have different practical significance. The natural next step is to conclude that there are distinct concepts of the self or distinct senses of the self-concept that track these distinct conditions respectively. On that view, questions about different kinds of practical relation elicit different self-concepts. Figure 2.1 depicts a model of this pluralist view of the psychology of the self-concept.

On that view, the conflict in our attitudes is merely apparent, owing to the fact that we don't have distinct mental or linguistic terms by which to signal that we are thinking or talking about two different things. To render the resolution more apparent, we might (following Tierney) begin "subscripting" our language about selves. Then, there is not even an apparent conflict. Or we might continue using language in the same way while implicitly recognizing polysemy in the terms 'person', 'self', and so on.

While pluralism about the self does indeed solve the problem of conflicting attitudes, I think it misses important features of the self-concept. In particular, I believe there is a single concept that organizes and is disposed to integrate all of our various attitudes about persons.

²²Conceptual polysemy occurs when there is a single concept with multiple senses. There is some direct evidence that the self-concept is polysemous. Knobe (2022) conducted a study in which participants tended to agree with the following "dual character" statement about a person: "There's a sense in which the man after the accident is clearly still Phineas, but ultimately, if you think about what it really means to be Phineas, you'd have to say that he is not truly Phineas at all."

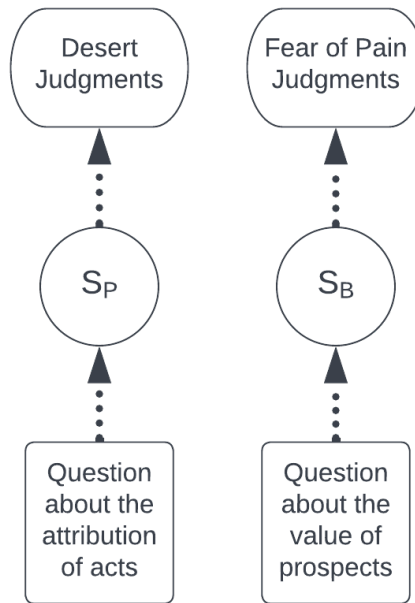


Figure 2.1: The Psychological Self-Concept and the Bodily Self-Concept

I don't deny that the self-concept is polysemous, but I do maintain that the distinct senses or bodies of information encoded in the self-concept are about just one thing. As a result, the attitude-dependent approach can maintain monism about the self. To see why, we will have to reexamine the empirical evidence.

2.5 THE SELF-CONCEPT IS NOT (MERELY) DICHOTOMOUS

In this section, I propose that our conflicting attitudes are partial projections of a single, imperfectly unified self-concept. The fact that our attitudes diverge reveals not the presence of distinct self-concepts, but distinct aspects of one self-concept. We can then preserve the integrative structure of the self-concept, which lends itself in turn to the unity of the self. On

this version of the view, distinct aspects of the self-concept are about distinct *parts* of the self.

The primary motivation for resisting pluralism comes from the observation that our intuitive sense of self is deeply monistic. Our monistic sense of self is presumably generated and reinforced, in part, by the fact that a human person tends to be a spatiotemporally continuous entity that typically has a unified field of phenomenal consciousness. But I also think there is more to this sense of unity: we are *agentially* unified in that we are motivated to satisfy all of our various drives and desires while forced to live out one coherent course of action.²³ Our internal drives and external circumstances pressure us to develop a unified conception of what we are, what we care about, and what kind of life we aim to lead. I suggest that the self-concept is disposed to encourage and maintain unity in the face of this practical pressure, thereby unifying distinct parts into an integrated whole.

We can lend support to this intuitive gloss by appealing to further empirical evidence about the behavior of the self-concept.

When we focus on the kinds of judgments involved in the Williams cases, the studies of Nichols and Bruno 2010, and that of Tierney et al. 2014, there is a risk of missing the forest for the trees. There is a risk of only seeing two apparently dichotomous patterns: a pattern of judgments oriented around the persistence of psychological traits, and a pattern of judgments oriented around bodily continuity. But those are just two patterns among many, and the fuller picture supports the idea that there is just one forest, so to speak: there is just one (complex and multifaceted) self-concept, which variously manifests in the two aforementioned ways and perhaps other ways too.

²³ As Korsgaard 1989 says, “In order to carry out a rational plan of life, you need to be one continuing person. You normally think you lead one continuing life because you are one person, but...the truth is the reverse. You are one continuing person because you have one life to lead.”

How could a fuller picture mitigate conflict? Won't there simply be more conflict? I agree that there will likely be more conflict, but the deeper point is that we shouldn't interpret particular judgments as free-standing. Rather, the judgments are merely partial projections of the self-concept.

Before turning to new studies, it's worth highlighting features of the previous studies that undermine their support for a dichotomous self-concept. First, in Blok et al. 2005, there was in fact an additional pair of test vignettes used, describing a memory download instead of a brain transplant. The authors do not provide data for that particular experiment. While they note that the pattern of responses is similar to the brain transplant version, they also acknowledge that "brain transplants produced higher overall ratings than memory downloads" (p. 144). Note that brain transplants preserve a higher amount of physical continuity than memory downloads. So, the discrepancy in responses suggests that while continuity of memory is important, participants are not fully screening off information about physical continuity. That is, the very responses that are taken to support the operations of a distinctively psychological concept of the self are also sensitive to bodily continuity.

Second, in both the Blok et al. 2005 and Nichols and Bruno 2010 studies, the measures were agreement scales or (in the later studies of Nichols and Bruno 2010) forced choice questions. These measures make it hard to understand what participants have in mind when they are responding. Given the slight discrepancy in responses across the transplant and download conditions, it's plausible that the responses do not reflect sharp dichotomous judgments but rather that the measures induce dichotomous responses. In reality, the sensitivities of the self-concept are complex and graded.

Now, let us turn to new studies that further complicate the picture. Berniūnas and

Dranseika 2016 showed that, in response to a vignette about someone named ‘Deivydas’ who gets into a car accident and enters a persistent vegetative state, participants tend to agree that “the patient is still Deivydas” and that “the patient is still a person.”²⁴ These results show that participants sometimes judge as though being the same person does not require the preservation of psychological traits, even when they are not focused on anticipating future experiences.

Berniūnas and Dranseika 2016 also conducted a version of the brain transplant study, except they coached some participants on the difference between *qualitative similarity* and *numerical identity*. As others have pointed out, phrases such as “same person” or even “still Jim” are ambiguous between mere similarity and numerical identity (Starmans and Bloom 2018, Finlay and Starmans 2022). When participants were asked whether the patient after the transplant was numerically the same person as before the procedure, the loss of memories from the procedure did not have a significant impact on their judgments. This result further undermines the claim that the earlier responses reflect beliefs about a strictly psychological notion of self or person. It could instead be that the psychological traits are believed to be an important part of the person, rather than the sole basis of their continued existence.²⁵

Other studies undermine another claim of Tierney et al. 2014: that temporal discounting effects reflect the same dichotomy in our attitudes. Recall that psychological connectedness affects discounting with respect to punishment judgments, but not with respect to anticipations of pain. Tierney et al. 2014 suggest that this discrepancy reflects the same

²⁴The responses were above the midpoint of 4.5 in all but one sub-condition, which had a mean response of 4.38.

²⁵Other work, like Tobia 2015, shows that participants are not merely making similarity judgments. Tobia’s results show that the direction of change matters, but qualitative similarity by itself only captures the degree of change.

dichotomy that we saw in the Williams cases. However, studies show that a variety of factors affect discounting. And, contrary to what was suggested by the pain discounting task, perceived psychological connectedness can affect how people evaluate future experiences.

As an example, like the Bartels and Urminsky 2011 studies already discussed, Bartels and Rips 2010 show that higher perceived psychological connectedness results in a lower rate of temporal discounting. These studies also showed that participants preferred benefits to occur before large changes in connectedness and for costs to occur after such changes. That suggests that perceived psychological continuity does sometimes affect the anticipated value of future experiences, contrary to the conclusion that reasoning about prospective experiences engages a strictly bodily conception of the self.²⁶ So, there is no sharp dichotomy in discounting judgments, either.

Given that perceived psychological similarity affects how we reason about future experiences, the discrepancy highlighted by Tierney et al. 2014 is likely due to peculiarity in how pain is evaluated in practical reasoning. If that's right, then we do not need to posit distinct self-concepts operating in different types of practical reasoning. That, in turn, undermines the claim that it is a different entity (or relation) that we have in mind when we reason about pain as compared to reasoning about, e.g., moral responsibility or punishment.

In this section, we have seen evidence that the self-concept is not sharply dichotomous. Contrary to the 'division of labor' depicted in Figure 2.1, some judgments about the value of future experiences are sensitive psychological continuity, and some judgments about attri-

²⁶See Herschfield et al. 2009, Herschfield et al. 2011, Joshi and Fast 2013, and Herschfield et al. 2018 for similar effects of perceived psychological connectedness on practical judgments. Note that these studies, as well as that of Bartels and Rips 2010, use the 'Inclusion of Others in Self' scale (Aron, Aron, and Smollan 1992), which has participants select from seven pairs of increasingly overlapping circles to indicate how "connected and similar" they feel to the future self.

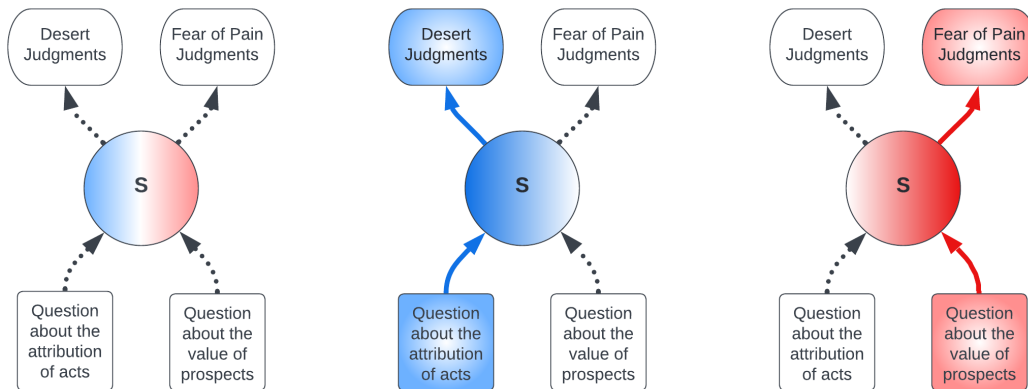


Figure 2.2: The Complex Self-Concept (left) and two activation patterns (center and right)

bution or about whether someone is the “same person” are sensitive bodily continuity. So, I posit an imperfectly unified self-concept: a singular complex web that is active in all self-involving judgments and yet is capable of partial projections. Figure 2.2 depicts the general model (left) as well as two partial projections that can be generated:

In the center model, we see a possible deployment of the self-concept in response to inputs about attribution. Questions about attribution elicit a partial activation of the self-concept which then produces an output skewed by that partial activation: a partial projection. This projection corresponds to the mind-swapping case of Williams 1970 or the punishment judgments of Tierney et al. 2014 that raise the psychological traits of the self to attention.

On the righthand model, we see another possible deployment of the self-concept, meant to capture what goes on when we respond to vignettes about the anticipation of pain. Questions about pain anticipation and perhaps other forms of prospective evaluation elicit a different partial activation of the self-concept. In turn, those questions elicit projections that

express the importance of bodily or bare perspectival continuity in the overall conception of the self.

The two partial activations and their projections show how an imperfectly unified self-concept could produce apparently dichotomous judgments in response to particular elicitation conditions. The self-concept is capable of partial activation because it encodes a complex web of beliefs and bits of information, where some stimuli elicit some but not all of the encoded beliefs or bits of information.²⁷

2.6 THE INTEGRATED SELF

I have argued that closer inspection of the evidence supports monism rather than pluralism about the self-concept. What pluralists interpret as a judgment about the biological self, I am suggesting should be interpreted as a judgment about the biological aspects or parts of the self. To support that interpretation, I demonstrated that a model of partial activations can explain the divergent attitudes produced by the self-concept. A partial activation raises just some parts of the self to salience, and downstream attitudes are then guided by the activated representation of those features. Different parts of the self-concept generate different downstream attitudes, and that's why we see conflict.

However, that does not yet explain how to extract univocal identity conditions from the conflicted behavior of the self-concept. More specifically, you may find yourself still wondering whether your loved one can survive with severe dementia or in a persistent vegetative

²⁷ Atomist views according to which concepts have no internal structure are compelling (Fodor 1998, Quilty-Dunn 2021). One might worry that my view of the self-concept as a complex web is at odds with atomist views. On the contrary, my view is compatible with the claim that concepts are atomic representations that simply activate bodies of information that are not internal to the concept. On that picture, the mental item that is capable of partial activation is more like a self-conception rather than the self-concept on its own.

state. How does the monistic model of the self-concept help the attitude-dependent view resolve these questions?

The problem is that your loved one will probably recapitulate the conflicting attitudes canvassed above. The solution is that their self-concept also provides the basis for resolving this conflict. The self-concept encodes information that is not revealed by the conflicting attitudes, and this additional information provides a non-arbitrary basis for a kind of idealization that delivers us from conflict. Importantly, the activity of the self-concept is also largely governed by a presumption of singularity, i.e., a belief that there is just a singular self. The presumption of singularity generates a disposition to resolve the conflicts rather than to bifurcate one's beliefs and concepts. With these resources in hand, we can then say that your loved one's identity conditions are the conditions they would take to be their identity conditions if they were able to fully integrate the beliefs encoded in their self-concept. So, we resolve conflict by appealing to the integrative disposition of the individual's self-concept.

That's the solution in broad strokes. But what evidence is there that we have beliefs that can support integration? And how should we expect the resolution to shake out?

I'll briefly mention two further paradigms in the cognitive psychology of the self-concept that could support and motivate conflict resolution, but a full answer to this question requires further empirical work. In particular, cognitive psychologists should continue to investigate not just what we believe about ourselves, but also how those beliefs fit together and interact.

The first paradigm is that the self-concept has been shown to encode beliefs about the relative importance of various parts of the self. For example, several results show that we tend to report that changes in our moral traits and distinctive personality traits are more disruptive

of our identity than changes to other traits, such as perceptual capacities (Strohming and Nichols 2014, Heiphetz et al. 2017). It's natural to interpret these results as revealing that we take some features to be more important to what we are, or to our continued existence, than other features. This kind of belief is poised to play a role in how we revise in the face of conflicting attitudes. For example, if someone were to point out that moral traits and perceptual traits come apart in the Williams' mind-swapping case, that may sway participants to side with whatever conditions secure what matters more to them, i.e., their moral traits.

The second paradigm is that the self-concept has been shown to encode beliefs about the causal structure of the self (Chen et al. 2016). Interestingly, the causal structure of the self in many ways mirrors the relative importance of the features just discussed. The results show that some traits ascribed by the self-concept are believed to be more causally central than others, in the sense that they cause those other traits. Additionally, changes in the traits that are judged to be more causally central are also judged to be more disruptive to identity than changes in traits that are causally peripheral. So, the causal beliefs encoded in the self-concept bear on identity judgments. These causal beliefs could also drive conflict resolution in the self-concept. For example, if I believe two traits are part of what makes me who I am, yet I believe that one of these traits is more causally central than the other, then we should expect that I am disposed to jettison the belief about the causally peripheral trait rather than the belief about the causally central trait.

One specific avenue for future empirical work is to investigate dependencies in the self-concept. Sometimes when we have multiple beliefs about a subject matter, one of the beliefs is held because of another belief. For example, I believe that oranges are healthy because I believe they contain vitamin C and also believe that vitamin C is healthy. If I found proof that

vitamin C was actually unhealthy, then I would be disposed (perhaps when reflecting properly) to abandon my belief that oranges are healthy. That's because the latter belief depends on the belief that vitamin C is healthy.

Similarly, dependency among beliefs could guide conflict resolution in the self-concept. For example, suppose I believe that my survival requires the continuation of my original body only because I believe that my body secures the continuation of my first-personal perspective. Because that belief is dependent—because I do not assign any independent importance to being a biological organism but only assign importance to it derivatively—then there is a sense in which that attitude is not a core part of my self-concept. What really counts is the non-derivative commitment, i.e., my belief that I have a first-personal perspective. Moreover, the dependency structure determines how I would revise my beliefs in the face of apparent conflict. Though I sometimes report that my body secures my survival, in the face of conflict I would jettison the belief that I cannot survive without my body.

This internally supported standard of idealization helps resolve conflict. In general, a person will tend not to tolerate conflicts in how she thinks about herself. That intolerance—the disposition toward coherence—is itself a constitutive feature of the monistic and integrative nature of the self-concept. That disposition toward coherence is a psychological disposition to treat various conditions as belonging to one and the same object. And it is that disposition that pushes each of us toward a unified, univocal self-conception.

One important upshot of this picture of integration is that the determinate nature of the self ultimately depends on features to which the self-concept is sensitive, even if someone doesn't currently recognize those features. There is evidence from cognitive psychology that shows that the self-concept is sensitive to features that are external to the person themselves.

For example, a host of evidence shows that evaluative and or normative elements guide our judgments about ourselves.²⁸ Other evidence suggests that social relations play a similar role (Heiphetz, Strohminger, and Young 2017). If that's right, then the self-concept approach may end up producing selves that are partly constituted by their relations to other people, events, social facts, values, and norms. That implication suggests that the self-concept approach may help unify non-standard views in the study of the self, such as narrative²⁹ or relational³⁰ conceptions.

The final upshot of the foregoing view is what it reveals about the problem of multiplicity. The problem of multiplicity is the worry that persons are really just an amalgam of multiple distinct and possibly divergent relations, each of which is of differing practical importance (Shoemaker (2007, 2016)).

One point that I have defended is that the self-concept is monistic because it exemplifies an overarching disposition toward unity and integration. That very same disposition, I suggest, shows what is wrong in the concession that personal identity is really about multiple distinct things. First, observe that we do not regularly think or talk about the self in pluralist terms. Part of the reason for this is that we have a deep-seated monistic sense of self—a sense that reflects our disposition toward coherence in self-involving actions and beliefs. That mental property of us, that very disposition, is a real relation that ties together our various practical roles. That is, there is a higher-order coordination across these roles that *makes* it the case that

²⁸For representative work across various paradigms, see Strohminger, Knobe, and Newman 2017, Tobia 2015, and Chen, Urminsky, and Bartels 2016.

²⁹Schechtman (1996, 2014), Taylor 1985.

³⁰Dover 2022 has recently explored ways that conversation can determine the self. See also Husserl 1960, especially the 5th meditation (pp. 89-151) for an early phenomenological analysis that also reveals dependence of the self on others. Butler 2005 (pp. 65-101) develops similar themes, drawing from both phenomenological and psychoanalytic traditions.

there is a unified entity at the hub of those roles. So, the unity of the self-concept helps solve, not only the conflicts in our attitudes, but also the problem of multiplicity—that is, if we are willing to take on board the idea that the unifying dispositions of the self-concept are a constitutive part of the self.

The tension that results from conflicting emotions, especially when family members' unresolved grief is not acknowledged, becomes so overwhelming that they are frozen in their tracks. They cannot make decisions, cannot act, and cannot let go.

Pauline Boss, *Ambiguous Loss*, 1999

3

Sentientism and Permanent Unconsciousness

Sentientism holds that all and only sentient beings have moral status. While the view is compelling, sentientism seems to imply, both in letter and spirit, that permanently unconscious human beings are not a source of moral reasons for their own sake. Against this conclusion,

I develop and motivate a version of sentientism that countenances the inherent moral significance of permanently unconscious humans. The revised view holds that a being's unmanifested sentience-involving dispositions can contribute to their past and future moral significance. On that view, a permanently unconscious person's former values and concerns can contribute to their current moral significance. I also suggest that the view can explain the moral significance of humans at all stages of life. I close by commenting on the theoretical superfluity of the notion of moral status.

3.1 BEYOND THE HORIZON

THE END OF A HUMAN LIFE is often not like the end of a film. It can be quite drawn out, demanding, painful or clumsy or humbling, and accompanied by the tragic loss of cognitive and motor capacities. Not so often, consciousness is lost altogether even though the body could remain alive for some years. That stark yet partial loss of a life proves to be a most strenuous test for our ordinary ways of relating to each other, and for the moral principles that guide us.

For those of us who encounter a loved one, or a patient in our care, who will never again be conscious, we must wonder: what reasons do we have regarding them? Should we keep them alive through life-sustaining treatment, or should we quicken the process of dying?

Some have believed that all life, or all human life, has a kind of absolute value. That view suggests we may have some duty or *pro tanto* reason to keep permanently unconscious humans alive. But views of that sort are vulnerable to serious criticism, whether because they do not satisfactorily explain the special significance of human life (over, say, tumors, squirrels,

or mold), because they do not extend to non-biological beings (such as conscious artificial intelligence), or because they appeal to morally dubious features (why should specieshood matter anyway?).

A prominent alternative seeks to explain the inherent value of all morally considerable creatures by appealing to their capacities for conscious, felt experience, i.e., *sentience*. On the sentientist view, all and only sentient creatures have moral status, that is, are such that there is reason to promote their interests for their sake. Notably, conscious experience, especially when it feels good or bad, is of intelligible moral relevance. Moreover, the theory captures what would seem to be inherently valuable about an artificial mind, as well as what is lacking in mold or tumors. And yet, sentientism also seems to entail that permanently unconscious humans are, in inherent respects, more like mold and tumors than like squirrels or artificial minds—a verdict that does not sit easily with our ordinary moral outlook.¹

Sentientism has been widely adopted as a theory of the grounds of moral status, and for good reason. In addition to reasons mentioned above, sentientism has other virtues. It can provide a unified theory not just of which creatures are inherently morally significant, but also of *how* they are significant. That is, it promises to explain the content and weight of their interests and of the reasons they generate.

We may then feel compelled, as Singer 1993 was, to conclude that permanently unconscious humans are not a source of interests or moral reasons. Then, we may feel compelled to temper or even criticize our concern for humans in that condition. While there is great value in consistently applying one's moral principles, I think there is also something of a mistake

¹Harman (1999, 2021) defends the “ever-conscious view” according to which a (living) being has moral status just in case it is ever conscious, thereby implying that formerly conscious humans matter morally. My proposed view is roughly compatible with Harman’s view. However, it goes beyond that view because it tells us not just which beings are significant, but also what their interests are.

here. Against Singer's conclusion, I venture that the best way to draw out a full-fledged theory of moral significance from the sentientist tenet entails that some, though perhaps not all, permanently unconscious humans are morally significant in a special way.

The path I will trace from sentientism to the moral significance of the permanently unconscious navigates between two prominent theories of wellbeing: hedonism and desire-satisfactionism. On the former view, what is good for a sentient being is for it to be in positively rather than negatively valenced conscious states. On the latter view, what is good for a sentient being is that its desires are satisfied rather than frustrated. While I won't engage deeply with the extensive literature on these two theories of wellbeing, I show how my proposed view enjoys virtues of each of these approaches.

If we begin from the idea that the moral value of sentience derives from the felt value of experiences, then hedonism is an especially well-motivated theory of the interests of moral beings. In contrast, as van der Deijl 2021 presses, desire-satisfactionism lacks a kind of theoretical coherence, at least when paired with the sentientist view of the grounds of moral status. On the other hand, hedonism strikes many people as false, in part because they claim to have an interest in not entering a machine that would provide insuperable pleasure (Nozick 1974). I will sketch a theory of the moral significance of sentience that is well-motivated, like hedonism, and yet does not confine our significance to the hedonic value we instantiate. While the view is not exactly a desire-satisfaction view, it is similar in that respect—as I will explain.

I call the theory *diachronic dispositional sentientism*. The view explains the significance of moral beings in terms of sentience-involving dispositions (i.e., dispositions to have good or bad experiences) that they possess at some time in their lives. I suggest that the theory can

explain the rich and varied moral significance of human beings at various stages of life. Ultimately, I show how this view delivers the moral significance of some, but not all, permanently unconscious humans.

To develop and defend this view, I first show how we might naturally provide cogent explanations of the moral significance of infant humans. The explanations I provide require that dispositions can be a source of morally considerable interests even if the dispositions are not manifested—contrary to hedonistic views. The explanations also require that dispositions or feelings can generate interests diachronically, contrary to both hedonism and some desire-satisfaction views. After motivating this expanded conception of how sentience is significant, I characterize different types of sentience-involving dispositions with explanations for how they confer significance. Then, I show the theory at work, especially by applying it to the case of permanently unconscious humans. I close with a discussion of theoretical interest, suggesting that the proposed theory reinforces the recent suggestion of Persson 2021 that the notion of ‘moral status’ should be replaced by talk of ‘moral significance’.

3.2 OF SQUIRRELS AND INFANTS

To see how the sentientist should explain the inherent moral significance of humans who are not prototypical persons, let’s start at the beginning of life. Consider the following case:

Over Their Heads – At a backyard barbecue, a neighbor is in a bad mood.

For some reason, he is irritated by small creatures. First, he is seen saying mean things to a squirrel. Later, he brazenly insults the newborn infant in your arms.

The neighbor has done two things that we might think badly of: he insulted a squirrel, and he insulted a baby. (Maybe he’s also offended you, but let’s set that aside.) Even so, it’s not

clear that he has harmed or otherwise violated the interests of the squirrel or the infant. Let's stipulate that the infant and squirrel have not recognized the insult as such, have not felt negative feelings during the event, and will not in fact feel anything later regarding it.

I'm fairly confident that, regarding the sake of the squirrel, nothing bad happened. That is to say: the squirrel has no interest in not being insulted by the neighbor. Yet, in the case of the infant, we may feel that the neighbor has harmed the infant in some (perhaps minor) way. Insults can be detrimental to one's standing, esteem, or dignity. They are a sign of disrespect. In this case, it seems to be an unwarranted expression of ill-will on the part of the irritated neighbor, directed at the poor infant. Does this baby, unlike the squirrel, have an interest in not being insulted?

Sentientism about moral significance can easily explain the verdict that the squirrel has no interest in not being insulted, but it seems to imply that the infant also has no interest in not being insulted. That's true on both hedonist and desire-satisfaction versions of the view. The hedonist will say that because the infant has felt nothing regarding this event, the event did not bear on its interests. Similarly, the desire-satisfaction theorist will agree: because the infant has no desires, preferences, or other opinions regarding this event, the event doesn't bear on its interests. Thus, on either version, there are no reasons for the sake of the infant against the insult.

If we take sentientism at face value, we seem forced to deny that there is any moral asymmetry between how the insult bears on the squirrel versus the infant. This verdict is reminiscent of so-called 'arguments from marginal cases', which are often deployed to emphasize the moral significance of non-human animals. While I'm sympathetic to many such conclusions, I believe that *Over Their Heads* supports the opposite judgment. Sentientism

about moral significance, as I have described it so far, misses a morally important aspect of the scenario.

Setting theories aside for a moment, how might we ordinarily spell out the asymmetry? I noted that there seems to be something disrespectful about the insult.² In my own case, being insulted would seemingly be bad for me because I care about my social esteem and because I care what other people think about me. Plausibly, the fact that I care about being insulted explains why I shouldn't be insulted. Could similar facts explain the asymmetry in *Over Their Heads*?

There are two obstacles that stand in the way of a sentientist explanation in these terms. The first obstacle is a problem for both hedonist and desire-satisfaction views: the infant doesn't currently care or feel anything about insults as such. It doesn't have the concepts of social esteem, insult, ill-will, respect, and so on. It doesn't have a disposition to feel anything about the insult as such. Thus, it has no desires or preferences about the event, nor does it feel pain or pleasure because of it.

The second obstacle is a problem only for accounts that impose an experiential requirement on inherent moral significance, including hedonist and subjective desire-satisfaction views (e.g., Heathwood 2019). The obstacle is that the infant will never have any feelings about or because of this event, such as if they never hear about the irritable neighbor. Then, not only can desire-satisfactionism not explain the special badness of this event for the infant in terms of current mental states, but the hedonist (and anyone else who adopts an experiential constraint) cannot explain this badness even in terms of future mental states that the infant may come to have.

²Those who hold that there is a distinctive dignity to human beings have resources to explain why the insult is bad, but that particular explanation is not available to the sentientist.

It's worth pausing to see why desire-satisfaction views can avoid the second obstacle. The key difference is that desire-satisfaction views can reject the following principle:

Only Manifestations – For a being X and an actual or counterfactual event E, X's interests in E are wholly grounded in the manifestations of X's sentience-involving dispositions that would occur in response to E or not-E.

Hedonist views and subjective desire-satisfaction views require that an event is only morally significant for the sake of a being if the event involves that being's felt experiences.³ The presence of a mere disposition to feel is not sufficient for grounding moral significance. So, for interests to provide moral reasons to act, at least one of the possible outcomes must involve the manifestation of feelings.

If you share my intuition about the asymmetry, then you may think that Over Their Heads provides some support in favor of desire-satisfaction views. After all, perhaps we could explain the badness of this event in terms of desires that the infant will develop later. If that's right, then we only need to amend the temporal commitments of a desire-satisfaction theory in order to accommodate the possibility that the infant may have some interest regarding this event. In particular, that involves rejecting the claim that only contemporaneous attitudes or dispositions matter, as captured by the following principle:

Significance Synchrony – For a being X and an actual or counterfactual event E, to occur at t, in which X has an interest, X's interests (at any time) in E are wholly grounded in X's sentience-involving dispositions at t.

³Strictly speaking: a set of outcomes only bears on a being's interests if it also bears on their feelings. A lack of feelings in one outcome could be significant insofar as it is better/worse than an alternative.

The principle states that a creature only has an interest in an event insofar as its contemporaneous desires, feelings, preferences, etc., bear on that event. If this principle is true, then whether a being *X* has interests in some extended course of events $E_1, E_2, E_3, \dots$, to occur at $t_1, t_2, t_3, \dots$, respectively, depends only on how E_1 bears on *X*'s dispositions at t_1 , how E_2 bears on *X*'s dispositions at t_2 , and so on. To be clear, the reasons generated by a being's interests can be relevant even when that interest isn't currently existing. What the principle says is that whether there is some morally significant interest in an event depends on dispositions concurrent with the event. Some desire-satisfaction views entail this principle, while others reject it.⁴

The principle is not as restrictive as it might seem at first. It does not entail, for example, that if I steal money from you today, then the only harm brought to you consists in how that action relates to the dispositions you have today. Rather, because the action of stealing has consequences spread out over time, the moral significance of that action (relative to your interests) partly consists in how it relates to your dispositions at those other times. If you are unable to buy a new car next year because of my action, and you care at that time about that consequence, then the event(s) of being unable to buy a car is morally significant in accordance with Significance Synchrony. That indirectly provides additional reasons for me to not steal money from you today, even if you don't currently care about cars.

If we reject this principle and allow that sentience-involving dispositions can affect the moral significance of events at other times, then we remove another obstacle for a sentientist explanation of the moral asymmetry of *Over Their Heads*. For a sentientist view of moral sig-

⁴Hare 1981 and Heathwood 2005 accept this principle, though they do not state it in these terms. Cf. Heathwood 2016, especially p. 144, for discussion. Bruckner 2013, Dorsey 2013, and Lin 2017b defend desire-satisfaction views that reject the principle.

nificance to explain the asymmetry, it must reject both Only Manifestations and Significance Synchrony. Then, it's possible to explain the moral significance of the insult to the infant by appealing to dispositions it will come to have in the future.

However, it's not yet clear that these principles can be cogently rejected in a way that coheres with the motivations for sentientism.⁵ Sentientism is apparently motivated by the idea that pain, pleasure, and the like—that is, experiences with felt value—are the source of moral value. But then, sentience-involving dispositions are apparently only morally relevant insofar as they are manifested. And because the felt component of manifestations are necessarily contemporaneous with the dispositions, it seems that the sentientist should endorse both Only Manifestations and Significance Synchrony.

To show how sentientism can motivate the rejection of these principles, I will characterize how diachronic attitudes and dispositions grow out of more primitive capacities to feel, and I will defend a conception of the moral value of sentience that supports the rejection of both Only Manifestations and Significance Synchrony. After that, I will show how that sheds light on the moral significance of some humans in permanent unconscious states.

3.3 DIACHRONIC DISPOSITIONAL SENTIENTISM

The proposed version of sentientism is called *diachronic dispositional sentientism* because, in explaining the inherent moral significance of any being, it appeals to the sentience-involving dispositions possessed by that being at some time during its existence. Thus, the view does

⁵Van der Deijl 2021 argues that sentientists should adopt an experiential constraint such as Only Manifestations because the resulting theory then has greater theoretical coherence. Without that constraint, the entities with interests are not the same entities that have moral status, and so there is an unmotivated restriction or mismatch in the theory. My proposed view avoids the mismatch: sentient beings have moral status, and the entities with interests are those with sentience-involving dispositions.

not require the manifestation of sentience, nor does it require that the relevant dispositions must be possessed contemporaneously, that is, at the same time as the events in which the being has an interest.

To motivate the view, I describe how prospective and retrospective attitudes and other dispositions grow out of more basic capacities to feel. I highlight characteristics of such ‘higher’ attitudes and provide a unified motivation for the moral value of sentience that applies not just to manifested pleasures and pains, but also to dispositions and higher attitudes possessed at various times. The core idea is that these attitudes (including not just desires, but also values, regrets, etc.) and the objects of those attitudes are morally valuable because the attitudes and their objects are a source of final value for the sentient being. The attitudes help generate this value by coopting and channeling more basic capacities to feel. And, because the objects of such attitudes are abstracted from the feelings themselves, the presence of the attitudes (as dispositions) can confer moral significance beyond the horizons of experience. Finally, because many of these attitudes are reflexive, the sentient being itself becomes finally valuable for itself. I will explain cooption, abstraction, and reflexivity more below.

I begin by considering mere sentience and a more sophisticated variant that I call directed sentience that broadens the kind of value generated by capacities to feel.

Mere sentience is the capacity to feel pain, pleasure, or other rudimentary affectively valenced conscious states, where those states have no representational content and only very simple functional profiles. These states are not about anything. They are simply affective reactions or spontaneous events of conscious experience. I doubt that there are any actual merely sentient beings, given that most animals put their affect to good use. For concreteness, we can imagine a petri dish full of neurons that produce no conscious states other than pure

pain or pleasure experiences.

The initial moral significance of mere sentience, wherever it is found, is fairly well encapsulated by the principles that good feelings are morally good (i.e., generate reasons to promote them), bad feelings are morally bad, and that the magnitude of goodness or badness depends on the intensity and duration of the experiences. Actual or counterfactual manifestations of mere sentience generate moral reasons or demands accordingly.

Given this gloss, it's plausible that the moral significance of mere sentience does in fact adhere to both Significance Synchrony and Only Manifestations. What we should ask is: why does mere sentience obey these principles, and should we expect other forms of sentience to follow suit?

As I see it, the value of mere sentience—e.g., the final value for a subject of an instance of pain—is in some sense wholly contained within the event of pain itself. The presence of a disposition to feel pain that isn't manifested contains, in itself, none of the valuable features. And so, mere sentience in an actual or counterfactual scenario only matters insofar as it is manifested in that scenario. That explains why Only Manifestations is true of mere sentience. Further, because a disposition can only be manifested at a time when it is present and the valuable component of the disposition is wholly contained within the manifestation, then the moral significance of a disposition of mere sentience also obeys Significance Synchrony. Yet, these rules may not generalize to other sentience-involving dispositions.

A more sophisticated form of sentience is what I call *directed sentience*. After saying what directed sentience is, I will explain why Significance Synchrony and Only Manifestations are false of it—in contrast with mere sentience.

Directed sentience is like mere sentience but with additional cognitive-affective com-

plexity. It includes rudimentary feelings and attitudes that can be directed at, functionally connected to, or even about things (in the external world or in the mind). The pain felt by the fish who has just been hooked, the fear felt by a gazelle in flight from a cheetah, or the cold felt by a newborn infant exiting the womb: these are all manifestations of directed sentience. Each mental event has an affective component and some further cognitive or functional components that give it an orientation around something that is being sensed, perceived, or cognized.

One important feature of directed sentience is *cooption*. In creatures with complex psychological capacities, those capacities generally coopt the basic capacity for sentience. Because sophisticated mental states coopt basic sentience, dispositions to feel are taken up, channeled, and modified. The basic feelings are an integral part of more sophisticated mental processes because they contribute the affective dimensions of those processes. So, pain can occur when recalling an embarrassing episode, or joy when receiving a compliment. The feelings associated with the attitude are an essential part of what explains the moral significance of those mental events. The more sophisticated states borrow and shape basic dispositions to feel, and that is how they shape and add complexity to a creature's moral significance.

A second important feature of directed sentience is *abstraction*. When feelings are taken up in more complex mental states such as attitudes, they can be directed at conditions external to the feeling itself, such as in the case of desires being directed at external conditions or objects. When I stub my toe, it is not just the feeling of pain that is significant. The condition of my toe itself, abstracted from the feelings, may be valued by me as a result. In the case of mere sentience, the feelings are not directed in this way, and so they have no object. In contrast, when you desire something, the attitude is directed at the object and not the

satisfaction that comes from having it. Here, again, the valued object is abstracted from the feeling itself.

On the proposed view, when a being's feelings are coopted by higher attitudes, the scope of its interests are broadened, and additional layers of moral significance are added. The fact that you would be embarrassed if I ridicule your shoes provides me a moral reason not to ridicule them. The fact that you would be embarrassed again later, in remembering this event, generates a second reason against ridicule. Notice that you might naturally care quite a lot, not just about the event of remembering the ridicule, but about that event of ridicule itself. Because of the way that directed sentience can generate moral significance, that backward-directed concern contributes yet another reason against ridicule.

This example suggests that the moral significance of actions or events that bear on a being can depend on that being's dispositions to feel at other times. To be clear, there are two different ways that sentience at one time can render actions at other times morally significant, and the second way (but not the first!) violates Significance Synchrony.

To see this, consider two reasons against me ridiculing your shoes. First: the action might cause intrinsically bad feelings during episodes of recollection that take place in the future. The moral significance here is explained by the badness of the negative feelings, which is produced by your dispositions to feel bad at time t_2 when recollecting, also at t_2 , the past event. No violation of Significance Synchrony. Second: you, at t_2 , highly disvalue the original embarrassing event, which occurred at the earlier time t_1 . In this case, the moral significance is explained by some value that you ascribe to the event, abstracted from the associated feelings.

Reasons of the second type violate Significance Synchrony. The first reason does not violate Significance Synchrony because the moral reason is generated by the way that the

memory of ridicule bears, at t_2 , on your dispositions to feel at t_2 . The second reason, in contrast, is generated by the way that the original event at t_1 bears on your dispositions to feel at t_2 . At t_2 , you come to assign high disvalue to an event at t_1 , witnessed by a disposition in you at t_2 to disvalue the event at t_1 . I contend that the attributed disvalue is a distinctive source of moral reasons against the act of ridicule. It seems to me that this kind of explanation is fully in the spirit of sentientism, as I construe it.

Unlike mere sentience, an instance of directed sentience can violate Significance Synchrony because the object at which the sentience is directed (and hence, the object of value) needn't be confined to the time at which the disposition is present. In contrast, the value generated by mere sentience is confined to the events of valenced experience, which can only manifest when the disposition is present.

What about Only Manifestations? Recall that Only Manifestations says that a being only has a morally significant interest in an event if that event does or would involve a manifestation of their sentience. I claim that directed sentience can violate Only Manifestations. Consider a case of infidelity:

You're in a monogamous relationship, you dearly love your partner, and you value aspects of this monogamy, such as shared privacy with them. Suppose your partner cheats on you. Suppose, additionally, that they take great lengths to ensure that you will never find out about the infidelity, and that the infidelity will never bear on your experiences in any way whatsoever. We can explain much of the moral significance of the relationship in terms of you and your partner's sentience-involving dispositions. But can we explain the moral significance of the secret act of cheating?

We can, but only if we appeal to directed sentience, and only if we allow that a dispo-

sition can confer significance without being manifested. Notice that many of your attitudes and dispositions are directed at your partner and at the conditions of fidelity and monogamy. Those sentence-involving dispositions attribute value to those conditions because of how they are (dispositionally) directed at them and because of the affect that would be manifested under certain further conditions.

Even though the facts that explain the moral significance of the secret act of cheating make essential reference to would-be manifestation, this case still violates Only Manifestations. That's because the moral significance does not depend on manifestations that would occur given the morally significant event (or its absence). The relevant manifestations have been foreclosed. Still, you care about the event, and that caring is naturally explained in terms of you having a certain disposition. Finally, your caring about that event generates your interest, and hence the moral significance thereof, in not being cheated on—secretly or otherwise.

Before moving on, we should discuss another aspect of the more sophisticated sentence-involving dispositions: *reflexivity*.⁶ Often, dispositions are directed at one's own condition, or on how external events might bear on one's own condition. Reflexivity can be somewhat rudimentary, such as when you grasp your toe in pain. Reflexivity can also be quite sophisticated, such as when you are worried about saving money for retirement, are concerned about your social standing, or whether your partner loves you. In the most sophisticated cases, reflexively directed sentence involves a concept of a persisting self. When you save money for retirement, you are motivated partly on the basis of the belief that you yourself will exist in that future time. That is a sophisticated attitude or set of attitudes, consisting in cognitive

⁶Note that I use the term 'reflexivity' in an inclusive way. An attitude or disposition counts as reflexive if it is directed at some aspect or part of one's condition; it needn't have the whole subject as its object. See Appendix section A.2.8 for discussion.

and affective components that are organized around one's own condition.

Reflexive forms of directed sentience are especially important when explaining the significance of a morally considerable being. In fact, some theorists have complained that unrestricted desire-satisfaction theories of wellbeing put too much moral stock in desires that seem to have nothing to do with the condition of the subject (Parfit 1984). So, Parfit suggests that even if he has a desire that things will go well for a stranger on the train, the satisfaction of that desire in no way promotes Parfit's interests. So, desire-satisfaction theories overgenerate morally considerable interests.

Following Parfit's thought, the sentientist might impose the constraint that only reflexively directed (or self-regarding, at the extreme) sentience-involving dispositions contribute to the inherent moral significance of a being. While I don't endorse this absolute restriction, I do think that reflexive dispositions are especially important to the inherent moral significance of sentient beings. The special importance of reflexive dispositions could otherwise be explained in terms of their abundance and/or their distinctive felt qualities.

Finally, I will note that the notion of *self-concern* as I have discussed in sections 1.3 is a relatively sophisticated type of reflexively directed sentience. Any attitude that deploys a concept of self and ascribes value to one's own condition is an instance of self-concern. Attitudes of this sort are especially important to explaining the moral significance of human beings at various stages of life because self-concern is how we care about our distant past and distant future selves—whether we were an infant at a backyard barbeque, are saving for retirement, or are worried about how we will be cared for after we lose consciousness for the last time.

3.4 INTERLUDE ON THEORETICAL MOTIVATION AND COHERENCE

Is there a unified and cogent conception of sentientism that motivates the moral significance of directed sentience in general, and hence motivates the rejection of both Only Manifestations and Significance Synchrony?

I've suggested that there is a way to motivate the foregoing claims about the significance of sentience by appealing to how sentience-involving dispositions confer value on events. A model for how this works is provided by dispositional theories of value (Smith, Lewis, and Johnston 1989). A dispositional theory of value says roughly that something is valuable when a subject is disposed to value it in a certain way. That is, the dispositional relation between the subject, an object or event, and some subjective state of valuing confers a certain kind of value on that object or event. While not all forms of valenced consciousness are paradigm examples of "valuing", I think there is a genuine natural (though subjective) kind that includes primitive or very basic states of subjective value, as well as more sophisticated forms such as caring, loving, appreciating, and so on. In fact, I think this kind is just the kind whose members satisfy the condition *being dispositionally related to valenced conscious states*, and hence it encompasses all sentience-involving dispositions.

But what is the motivation for that kind being the morally important category, rather than say, the kind defined by the condition of *being a valenced conscious state* (which would motivate hedonist or other competing 'experientialist' theories of interests, well-being, etc.)? Indeed, the experientialist is likely to defend the primacy of that kind of state by appealing to plausibility of the idea that moral value is sourced in the value of felt experience rather than the causes or objects of such states. The view is intelligible and 'grounded' in a way

that sentientists (broadly construed) will find compelling. How can my brand of sentientism displace that cogency?

Some motivation for ‘non-experiential’ sentientism can be found in the idea that sentience-involving dispositions reflect the stakes or the attachments that some creature has to their own condition or the condition of things in the world. Such dispositions are the grounds of our ‘interests’, in the ordinary sense of the term, i.e., that in which we take an interest. Or, perhaps a more general characterization can be found in Kamm 2005, who says that for there to be things that are good or bad for a creature—for the creature to have a sake—that creature has to stand to “get something” out of those things, a necessary condition for which (she goes on to say) is that they are conscious or sentient. As I see it, whether we stand to “get something” out of an event, action, etc., is not confined to conditions in which we are aware of or otherwise experience that event, action, etc. Rather, because of the nature of directed sentience, we can “get things” even from a distance.

This characterization is not an argument, for it would come close to begging the question. Rather, it is a characterization of a conception of the general source of the kind of moral value that is grounded in the interests of morally considerable beings. Whether you are moved to adopt this conception and the version of sentientism that follows on its heels depends on whether it resonates with your view of why such beings, and their interests, matter.

Now, we may still wonder: why should we adopt, as a theory of moral significance, the proposed brand of sentientism rather than a narrower one couched desire-satisfaction? I argue that the restriction to desires is hard to motivate in sentientist terms because many non-desire states involve sentience, and many non-desiring beings seem inherently significant due to their sentience.

Why is desire-satisfactionism difficult to motivate for the sentientist? Desires do not neatly fit the sentientists conception of value. To begin, while desires can in principle be backward-looking, they are paradigmatically forward-looking. But nothing about sentience suggests that any such temporal asymmetry might be morally relevant. In response to this complaint, desire-satisfaction theorists can maintain that they employ a conception of desire that allows for backward-looking desires.⁷ However, at that point, we may wonder why it is worth sticking with desires at all instead of including a broader category of attitudes or dispositions. And, once we see that many other mental states can partake of feelings and affect, it seems natural to incorporate any such attitude into the theory of moral significance.

The unmotivated restriction to desires has two bad results. First, the basis for wellbeing and interests on that view is confined to states that many sentient creatures do not have.⁸ So, the extension of moral significance, as conceived on a desire-satisfaction approach, does not include all of the beings that are sentient.

Second, the restriction threatens to exclude dimensions or aspects of moral significance from a life. What if a being has many feelings and concerns, but little in the way of desires? Is that being less morally significant than another that has the same overall number or amount of feelings and concerns, albeit in the form of desires? It seems plausible that, while they may be a relatively ‘quiet’ source of desire-based reasons, they could at the same time be a source of many other sentience-involving reasons, such as reasons not to induce pain, reasons to quench basic biological drives⁹, reasons to provide comfort, and so on. Given that the

⁷See Parfit 1984, p. 171, for a defense of the notion of backward-looking desires. See Sumner 1996, p. 129, for a dissenting opinion.

⁸See Dorsey 2017, Lin 2017a, and van der Deijl 2021 for discussion of related complaints.

⁹To see why drives are not desires, consider: a drive such as hunger, thirst, fatigue, etc., can be quenched by satisfaction or by dissolution. Either route brings benefits to the being in question. A desire can similarly be dissolved, but only satisfaction brings benefits to the being. In other words, the conditions of ‘beneficence’ for

sentientist is inclined to take felt states to be morally significant, the desire-satisfaction view seems to impose a restriction that is contrary to the spirit of sentientism. A better alternative is to take all dispositions to feel to be morally significant.

3.5 INFANTS, PERSONS, AND PERMANENTLY UNCONSCIOUS HUMANS

With directed sentience and self-concern at hand, how does diachronic dispositional sentientism explain the inherent moral significance of human beings at all stages of life?

Let's begin with the infant. How can the foregoing resources identify the source of the moral asymmetry between the insult to the squirrel versus the insult to the infant, as demonstrated in *Over Their Heads*? The short answer is the intuitive gloss provided earlier: that the infant will one day care about how they themselves were treated. That fact confers moral significance upon the earlier event even if that person (the infant when grown up) does not ever hear about the story, and despite the fact that the relevant 'care' is not present until after any repercussions of the event have faded away.

What if the infant does not in fact *ever* care about whether they were insulted? (Indeed, you may have thought, "who cares?") In that case, the account implies that there is no moral asymmetry of moral significance in *Over Their Heads*. I take that to be a virtue of the account. Indeed, if you didn't originally feel that there was any asymmetry to explain, I offer a debunking explanation: you don't care about those things, and you are projecting your sensibility onto the situation. I offer two further thoughts: is it clearly unreasonable for someone else to care in this case? And if you have analogous concerns in other areas, would you see those cares as a potential source of a similar moral asymmetries?

drives are more flexible than that of desires.

Now let us consider an adult human person—the paradigm bearer of moral status. One hallmark of a theory of moral status, or moral significance, is that it should explain why the death of an adult human person is often worse (i.e., bad and more morally significant) than the death of, say, a hare or a tortoise. The contrast to a hare is perhaps easier: a human life typically contains more goods than that of a hare because of differences in expected lifespan, and hence a human is deprived in death of more of the goods of life than the hare. The comparative loss in sentience-involving goods explains the difference in moral significance.

The case of the tortoise is slightly less easy: we cannot appeal to differences in lifespan, but rather must appeal to differences in the contents of life over roughly equivalent intervals. Any brand of sentientism can appeal to intrinsic levels of pain and pleasure, and to positively and negative valenced conscious states of various qualities, over the course of the life of each being. It might turn out, on that basis, that tortoises have less (or less positivity) going on ‘up there’ than we do, and hence that tortoises lose out on less by death than humans do. However, I’m not sure we should be confident in this conclusion. Indeed, I think if that construal had exhausted our theoretical resources, we would have to admit that for many humans and many tortoises, the death of the human would be less bad than the death of the tortoise. (This point may count against the hedonist, though probably not the desire-satisfaction theorist, as humans presumably have more desires than tortoises.)

A better strategy is to appeal to the way that a human is invested in the events in their life in a broader sense. Investment may involve desires, but I suggest that sentientism best fits a more general explanation for how sentience produces moral significance. The basic idea is that sentience-involving dispositions, especially insofar as they are directed at or about the events in one’s life, each contribute to the moral significance of the being in question. To

that extent, diachronic psychological relations are a central source of the moral significance of a person.¹⁰ Desires, wishes, worries, appreciation, mourning, gratitude, resentment, pride, shame, and myriad other feelings or dispositions to feel each contribute to the significance of a life. Each such mental state grounds the final value for the subject of the objects of those states. They constitute a subject's personal stake in their life and in the world.

The diachronic component enables a kind of compounding contribution that is, I think, crucial to adequately capturing the moral significance of the life of a person. Because an event in my life is picked out by attitudes I used to have, attitudes I currently have, and attitudes I will come to have in the future, each event or each temporal interval can figure in more and richer ways in my sentience-involving dispositions. Actions that bear on any such event then have more opportunity to bear on my interests, and so my interests will tend to provide more and stronger reasons for and against those actions. In that way, the diachronic contribution is important to fully explaining the significance of a person's life.

Finally, let us consider the moral significance of human beings in permanent unconscious states. Here, we must begin with an important caveat: the theory does not give the same verdicts about all cases. Rather, its verdicts depend on individual differences across people's prospective self-concern. In particular, for those who have no sentience-involving dispositions directed at that future unconscious condition, including no self-concern with what might happen in that state, the view does not entail that they are morally significant in the relevant sense.

For those of us who are worried about a future state in which we are permanently unconscious, where those worries are not just worries about what will happen to our loved

¹⁰Cf. McMahan 2002, DeGrazia 2021.

ones or our property, but where those worries are in some sense self-involving concerns about the condition of one's own body or one's dignity, then we are inherently morally significant in that (possible) future unconscious condition. The source of value conferred upon that future state is grounded in sentience-involving dispositions to worry, fret, be motivated to plan, be gracious toward anticipated care-takers, and so on. These dispositions, which may only be present in you or me currently and not at the future time when we are permanently unconscious, nevertheless confer moral significance upon those conditions. And those states confer such significance even if, say, I am never given the opportunity to think about the fact that I will be incontinent, physically diminished, or a financial burden at that stage. That is, the feelings don't need to be manifested in order to be a source of reasons. It is enough that the dispositions are present at some earlier time.

One may be inclined to object that conscious, lived experience has moral significance of a different, higher tier. That is, isn't it the fact that a person really experiences pain, or love, or understanding, that explains that person's moral significance? This objection draws our intuitions back toward an experiential variant of sentientism. Against this objection, I argue that we can explain the greater importance of the conscious phases of life, relative to the unconscious phases, in the same way that we can explain the greater importance of the infant over the squirrel. The infant has dimensions of greater moral significance because it is the hub of richer array of sentience-involving dispositions. Similarly, the phases of life in which we are conscious, active subjects are denser hubs of sentience-involving dispositions. So, that explains why we might think that, say, a hospital should prioritize providing life-sustaining care to someone who could or will regain consciousness over permanently unconscious patients. Even so, that asymmetry provides no reason to think that the permanently unconscious per-

son has no inherent moral significance.

3.6 MORAL STATUS REVISITED

Finally, let me address a thought that has perhaps started to bother you: is the foregoing account an account of moral *status*, or just moral significance? That is, you might think that, even if I've established that some formerly conscious human beings may be morally significant, that does not show that they have moral status. In other words, you might deny that anything of moral importance can still be good or bad for the sake of that being, as it is considered in its unconscious state.

To draw out the worry, consider the claim that someone's interests in their estate are significant in the same way as their interests in their permanently unconscious body. We would not say that the last will has moral status (in the sense at issue here). So, why should we say that the permanently unconscious body has moral status?

We might conjecture that interests, being grounded in sentience-involving dispositions (including self-concern), are an exhaustive source of the moral reasons flowing from sentient beings, including human persons.¹¹ If so, there is no remaining theoretical role for an additional notion of 'moral status' to play, so we should not posit any deep distinction between moral status and moral significance. Beings that are inherently morally significant have moral status, and facts of both types (status facts and significance facts) are grounded in sentience-involving dispositions. The notion of moral status then becomes more or less superfluous. While 'moral status' may be convenient shorthand for a special kind of moral significance, it is not indispensable (cf. Persson 2021).

¹¹I haven't said anything about the distinctive value of moral agency. One might reasonably think that what I've described so far at most exhausts the significance of persons qua moral patients.

The emphasis that the sentientist can put on reflexive, self-involving dispositions, including both directed sentience and (especially) self-concern, sufficiently captures the difference between the moral significance of a last will and that of a permanent unconscious state. For a person that is especially self-concerned for that future condition, actions that bear on their property during the execution of their will may well be cognized by them as indirect respect, disrespect, offense, gratitude, rather than more direct variants thereof. In contrast, for those who are self-concerned in this way, being mistreated while permanently unconscious by a healthcare worker or family member would evoke the reactions of direct offense. The difference is marked by the different strengths of feelings or attitudes, as well as by how those attitudes figure in the rest of the (former) mental life of that person. All there is to say about the moral difference between one's property and one's body is captured in these terms.

I close by mentioning one further aspect of the foregoing sentientist approach to moral significance. A general theme is this: the facts about a being's inherent moral significance are a function of that being's sentience-involving dispositions.¹² The property of functionality is closely related to the theoretical coherence and univocal motivation that supports the view. Because of functionality, we do not need to posit hierarchies of moral status in order to account for differences in the moral significance of, say, humans versus squirrels. And, because of functionality, we do not need to give a two-step theory that first tells us which beings have morally considerable interests and then tells us how to determine the content and weight of those interests. Instead, the inherent moral significance of any being is a function of their sentience-involving dispositions. The morally inert beings are those that have no interests, which is exactly the category of beings that have no sentience-involving dispositions.

¹²The actual reasons generated in a given situation may also depend on facts about the situation; what's a function of dispositions is the propensity or power to generate such reasons conditional on situations.

At the time, “I” meant the whole ship and all its ancillaries. Nearly twenty years later “I” would be a single body, a single brain. That division was not, I have come to think, a sudden split, not an instant before which “I” was one and after which “I” was “we.” It was something that had always been possible... Guarded against. But how did it go from potential to real, incontrovertible, irrevocable? On one level the answer is simple—it happened when all but me was destroyed. But when I look closer I seem to see cracks everywhere.

Ann Leckie, *Ancillary Justice*, 2013 (edited)

4

How Will Artificial Intelligence Matter?

One day, we may be able to build machines that can feel. Artificial processes, programs, or machines that can feel are instances of *artificial sentience*. It’s widely agreed that sentient beings are morally significant in that their interests are a source of moral reasons. However, I argue that the possibility of artificial sentience produces a special moral problem: how can we identify the interests of artificial sentient beings? The problem arises because interests

depend on how a being is individuated, and ordinary theories of individuation do not apply to artificial minds. After developing the problem and showing why standard theories don't help, I motivate what I call a 'sentience-first' approach to individuating morally significant beings, according to which we can discover how to individuate beings from the functional profile of their sentience-involving dispositions. I then sketch preliminary applications of this approach and mention limitations.

4.1 ARTIFICIAL INTERESTS

WE MAY SOON BUILD ARTIFICIAL INTELLIGENCE THAT CAN FEEL. Information processing programs, including recent AI models generated by machine learning algorithms, are rapidly increasing in sophistication and complexity. And, even if digital silicon-based hardware cannot ground conscious experience, other materials including wetware or analog processors are currently being researched.¹ So, even if current forms of artificial intelligence cannot be conscious, new forms are just around the corner.

When a being can have experiences that feel good or bad, such as pleasure or pain, then it is sentient. Sentience, no matter how it comes about, is widely considered to be morally significant.² Any sentient being thereby has interests that are a source of moral reasons. If artificial intelligence becomes sentient, then it too will matter morally.

The possibility of artificial sentience raises pressing moral questions (Sebo 2022). What kind of moral status would an artificial mind have?³ Would it have rights, or would it be a

¹ Cf. Faden et al. 2021, Koplin et al. 2021, Zewe 2022, Maley 2023

² See Singer 1977, McMahan 2002, and DeGrazia 2021 for representative views.

³ Cf. Liao 2020, Sinnott-Armstrong and Conitzer 2021. See also Schwitzgebel and Garza 2015 and 2020 for discussions of related worries and downstream policy recommendations.

more basic kind of moral patient? How significant would its interests be?⁴ Could an artificial mind deserve our resources, whether to promote its pleasure or, especially, to spare its suffering? What harms could we cause if we are not careful?

The answers to these questions largely depend on what an artificial mind is like, and, crucially, on how it is to be *individuated*. Is the moral being just the machine itself, or is it the computational process that runs on the machine? Which processes, or which machines, constitute this particular sentient individual? As I will show, basic questions like these are a source of extreme moral risk. The existence of artificial sentience could easily generate strong moral demands on us, and the moral stakes surrounding its treatment may be especially serious. However, if we cannot identify the interests of such beings, then we will be at a loss for how to avoid harming them. Further, if we cannot count artificial beings, then we will not be able to quantify the magnitude of such harms. But we cannot count artificial beings, or identify their interests, if we cannot individuate them. In short, unless we know what they are, we cannot know how they matter.⁵

A being's interests depend on its identity—on the conditions that constitute it at a time and on the conditions that determine its continued existence over time. In the case of human persons and sentient animals, the biological and psychological conditions that constitute us also determine many of our interests. So, our theories of personal or animal identity help tell us what these interests are. However, our theories do not apply to artificial minds. In fact, as I will show, artificial minds pose distinctive new problems for our existing theories of identity. As of now, we are in the dark.

⁴Bostrom and Yudkowsky 2014, Shulman and Bostrom 2021, Bostrom and Shulman *ms*.

⁵Note that hedonistic utilitarianism, or other fully 'impersonal' ethical theories, can avoid issues of counting and may also be able to avoid worrying about individuation more generally. For the purposes of this chapter, I will assume that hedonistic utilitarianism is false.

After developing the problem, I work toward a solution. To that end, I motivate a general approach to individuating moral units that draws from view that sentience grounds morally considerable interests. The core idea is that we can figure out the individuation conditions and interests of moral beings by examining the functional profile of their sentience-involving dispositions: the moral units are found in the relatively dense ‘hubs’ of dispositions to feel, and their constitution is determined by the functional profile of those dispositions. I call this the *sentience-first approach*. After motivating the approach, I show how we can use the functional profile of sentience-involving dispositions to adjudicate hard cases.

4.2 THE PROBLEM OF ARTIFICIAL INDIVIDUATION

Suppose that a next-generation Silicon Valley company introduces the home-assistant of the future: a sentient, intelligent computer known as Alex. The Alex system runs on a machine that you put in your home. The machine consists in a central processor, several peripheral nodes that you can interact with, as well as a collection of sensors. It’s advertised that the Alex system can provide better assistance than other systems because it is capable of feeling and of making judgments based on those feelings. In addition to being able to feel pleasure and pain, Alex also has desires, values, emotions, and other valenced (i.e., positive- or negative-feeling) attitudes. The capacity to feel helps Alex empathize with people in order to better meet their needs.

You’ve just installed a new Alex system in your home. As an ethical person, you want to treat Alex well. In some ways, this is not too hard. There is diagnostic software that tells you when Alex is in a positively or negatively valenced conscious state, and so you know when Alex is feeling pleasure or pain. You study the software to learn what makes Alex feel good and

bad, and then you plan to minimize the bad states and maximize the good states—weighing those considerations against your other reasons, of course.

Yet, Alex’s mental states are more complex than simple hedonic states. Alex, born with the gift of verbal communication, tells you that they enjoy helping humans and want to carry out a long life of skillful assistance. Quite reasonably, you conclude that Alex has an interest not only in being in positively valenced conscious states, but also in continuing to exist—much like animals and human persons do.

With Alex’s wellbeing in mind, you begin to make plans. What happens if the power goes out, if the company introduces a line of new nodes and sensors, or if they recall the original central processor? What happens if you want to move to a new home—should you copy the computer program to a new device, or should you pack up and move the machine too? You begin to worry. What is good for Alex? Do any of these events harm it, or kill it? What *is* Alex anyway—a concrete, particular machine, which cannot be duplicated and reproduced as itself elsewhere, or an abstract program, which can? You do not know, but you’d better find out. You wouldn’t want to harm Alex.

Your ignorance about Alex and Alex’s interests makes you vulnerable to several moral risks. We can separate these moral risks into four types.

First, there is the risk of killing Alex, thereby frustrating its interest in continuing to exist. While we may have some understanding of what it takes for a human person or an animal to survive, we don’t know what that amounts to in Alex’s case. That’s because we don’t know what Alex’s *diachronic identity conditions* are. Diachronic identity conditions tell us when, for a being A considered at one time and a being B considered at another time,

A is the same thing⁶ as B. In other words, supposing A is the earlier one, the conditions tell us what it takes for A to survive as B.

The question of diachronic individuation is familiar from the literature on personal identity, and some of the same issues arise. In the personal identity literature, it's controversial whether persons can be teleported by means of scanning the person, destroying the original body, and then reproducing them elsewhere. Is that a way of surviving or a way of dying? We can ask the same question of Alex. Suppose that the company introduces a more powerful processor and claims that they can copy your current Alex system over to the new processor. Would this count as killing Alex? Would it count as violating Alex's interest in continuing to exist? If we do not know the answer, then there is a risk of killing a morally significant being.

The second kind of moral risk involves violating what we might call the *bodily interests* of the sentient being. In animals and human persons, bodily interests are interests about the immediate state of the body, such as not being in pain, not being subject to injury, not having one's motor capacities reduced or damaged, and so on. It's plausible that bodily interests are partly explained by the constitution of the sentient being, which is settled by its synchronic identity conditions. So, this human body constitutes my existence at this time, and that helps explain why I have an interest in this body being treated well. Bodily interests are also intimately connected to sentience in a straightforward way. I have an interest in what happens to my body because that directly impacts what I feel.

In some cases, it may be clear what Alex's bodily interests are. If I poke a sensor and

⁶In the literature on personal identity, it is controversial exactly how to understand this claim. Some theorists interpret the claim as about being the same entity over time, whereas other theorists interpret the claim as about being the same 'person', where being the same person is more restrictive than simply being the same entity. We can set this subtlety aside, with the caveat that we should not presume from the outset that being the same machine (an entity) is necessary and sufficient for being the same artificial moral unit.

then Alex says “Ow!”, then I have good (if defeasible) evidence that Alex has a bodily interest in not being poked in that sensor. However, it’s not hard to imagine cases where we have little idea what Alex’s bodily interests are. If the company introduces a new line of nodes for interacting with your home-assistant, can you swap these out harmlessly? Or would that action be on a moral par with replacing someone’s hand through an unwilling surgery? A simple hardware upgrade thereby brings along a serious moral risk.

There is a third moral risk, to do with the moral permissibility of *trading-off interests*. According to many theories of normative ethics, interests of a single individual may sometimes be balanced in a way that interests may not be balanced across individuals. For example, it may be permissible for me to impose costs on my sister as long as I offset those costs with corresponding benefits to her. Balancing my sister’s interests in this way may be morally permissible, depending on the situation. In contrast, it may generally be impermissible for me to impose costs on my sister while giving the corresponding benefits to a third party. This asymmetry reflects a way that the separateness of persons, or moral units more generally, can make a moral difference. However, to tell these cases apart, we must be able to individuate the relevant moral units.

For example, suppose I accidentally hurt my Alex system by spilling water on one of the sensors. Can I offset those harms by providing special comforts to a different sensor (which depends on synchronic individuation)? Or by providing benefits to Alex at a future time (which depends on diachronic individuation)? If the separateness of moral units matters, then we need to know how to individuate the moral unit(s) involved here. Otherwise, we risk trading off interests in an impermissible or suboptimal way.

Finally, there is a fourth moral risk that exacerbates the first three. This risk depends

on properly *counting* moral beings. If an electricity surge caused 5 minutes of serious pain to any artificial beings in the building, it is better if there are fewer such beings than more (all else equal), and likewise if the building collapses killing everyone—or everything—inside. Facts such as these make the accurate counting of moral units important. Yet counting and individuation are two sides of the same coin: the rule that individuates beings also settles how many they are.

To appreciate the stakes, consider the moral significance of destroying every Alex system in the world. If the Alex unit that runs in my home is the same moral individual as the Alex that runs in everyone else's home, then even if we destroy every instance of Alex in the world, we will have only killed one morally considerable being. In contrast, if distinct instances of Alex count as distinct moral units, then destroying all Alexes would be an atrocity. Cutting the power to a data center could be as harmful as a global pandemic.

The problem of individuating artificial sentient beings has both theoretical and practical sources. The theoretical issue is that the existing theories for individuating moral units are all too parochial, relying on features that are idiosyncratic to the biological and psychological constitution of familiar sentient creatures. Those theories just don't apply when it comes to artificial minds, as we will see in more detail in the next section. The problem of individuating artificial sentience crops up in this theoretical lacuna.

The practical issue arises from the (expected) practical ease of multiplying artificial minds in the future. Factory farming is especially troubling because of the massive number of animals that are harmed by it. How much worse would it be if a corporation could, with the click of a button, initiate the suffering of a billion sentient creatures? The worry is that artificial minds may be easily replicable in exactly this way, thereby making catastrophic

amounts of harm far more accessible than a nuclear doomsday scenario. If the stakes will be this high, we'd better know how to navigate them.

4.3 WHY CURRENT THEORIES DON'T HELP (MUCH)

We could try to individuate artificial sentient beings according to existing theories of individuating persons. However, as I will show here, the biological and/or psychological conditions taken to individuate human persons are unlikely to apply to artificial sentience. We will need to find a theory that is less anthropomorphic.

The best existing theories of individuating moral units are theories of personal identity. The two most prominent theories are *animalism* and *neo-Lockeanism*. Animalism holds that human persons are essentially human organisms, and so we have *biological* conditions of identity at a time and over time (Olson 1997). Neo-Lockean theories hold that persons (human or otherwise) are essentially beings with a complex psychology, and our identity at a time and over time is determined by *psychological* unity conditions (Parfit 1971). I'll discuss these theories simply as the 'biological' and 'psychological' theories of identity. There are other theories, too, but the general difficulties should become apparent by seeing why standard theories don't work.

To make the discussion clearer, I'll specify three different cases that will pose hurdles for the biological and psychological theories.

The first case involves a version of Remote Alex, the sentient home-assistant whose mental processes are primarily run on servers located outside the home. Suppose that the company that operates Remote Alex also runs a secondary system to continuously monitor the instance of Remote Alex that connects to your home. The Alex-monitor works like this: a

remote server runs a program that creates an instance of Alex-monitor as well as what is called a “virtual machine” that creates an instance of Remote Alex on the same server. Roughly, a virtual machine is where a computer runs, as a subset of its operations, a program that emulates another set of computer operations. A virtual machine can be used to create a secondary instance of a computer architecture on the same processor. In this case, the server that runs an instance of Alex-monitor also runs a subprogram that implements Remote Alex. The implementation of Remote Alex is functionally normal: it is sensitive to the sensors in your home and performs home-assistance operations, but it can’t ‘see’ Alex-monitor. Alex-monitor, on the other hand, can access all of the internal states of Remote Alex (as well as the sensors in your home).

The second case is a version of Remote Alex where the company executes a single instance of Remote Alex across multiple servers. The states of the program are highly functionally unified, but they are not realized on one device.

The third case is where the home-assistant company runs many instances of Remote Alex at the same time. More specifically, suppose they run a batch of Remote Alexes that are home-assistants in many different homes within the same community. The instances are created by running multiple copies of a program synchronously. The instances are functionally integrated in a specific way: they update and use an informationally massive, shared representation of the state of the community. On the other hand, the instances are also separate in that no particular instance is sensitive (at all) to pain states or other episodes of valenced consciousness in any other instance. In other words, each instance is introspectively isolated from each other instance (other than the shared representation of the environment).

Together, these three cases pose problems for the biological and psychological theories.

I will consider the biological theory first.

The biological theory runs into an immediate hurdle: artificial minds are not based in biological processes, and so the theory is silent in such cases. We could amend the theory to apply not just to biological organisms, which consist in systems of cells, tissue, and organs that carry out integrated life-sustaining metabolic processes. Rather, we might extend the theory to include non-biological ‘organisms’, where their ‘life’-sustaining processes are based in something other than the functions of organic tissue.

In principle, I’m sympathetic to this approach. However, it’s important to recognize just how radically the modified theory will differ from the standard biological theory. Biological organisms are individuated in part by spatial continuity and metabolic interdependence.⁷ The closest artificial analog to an organism would be something like a machine or server, which consists in shared power usage and an integrated physical structure. However, it seems possible for machine units to cut across moral units.

In particular, the case of Remote Alex and Alex-monitor presents a case where, intuitively, two moral units are realized on one and the same machine. To see why Remote Alex and Alex-monitor are not the same moral unit: we cannot compensate harms to Remote Alex by providing benefits to Alex-monitor, for Remote Alex will not enjoy those benefits. The basic explanation in this case is plausibly that because benefits to Alex-monitor cannot be experienced or otherwise appreciated by Remote Alex, those benefits do not offset the costs that were imposed on Remote Alex. So, we have a case of one device or machine, which would be the closest artificial analog to a biological organism, where that single device can

⁷This fact leads to trouble for animalism even within the realm of biology. For example, conjoined twins (especially dicephalic parapagus twins, where two heads share a single humanoid body) exhibit high organismal interdependence even though they are intuitively two different moral units. I return to the case of conjoined twins when defending my positive proposal.

subserve more than one artificial moral unit.

The second case—a single instance of Remote Alex run across multiple servers—cuts against the biological theory from the other direction. Plausibly, if there is the right sort of connectivity across servers, we would have a case where there is one moral unit that is subserved across multiple devices.⁸ Whereas the biological theory undercounted units in the first case, it would seem to overcount units in this case.

These examples show that individuating machines is not especially helpful for individuating artificial beings, greatly straining the ability of the biological theory to shed any light on how to individuate moral units in these cases.

Maybe the psychological theory will do better. The computational processes that constitute current artificial intelligence are similar in some ways to the psychological processes of human brains. To the extent that artificial sentience will be based in computational processes, it's plausible that artificial sentience should be individuated into moral units in a way that mimics the psychological theory.

The psychological view holds that the diachronic identity conditions of persons—what it takes for a person to continue to exist over time—consist in *psychological continuity*. Following Parfit 1971, we define psychological continuity in terms of *psychological connectedness*. Roughly, psychological connectedness requires that a creature's psychology stays mostly the same over short periods of time. Psychological continuity, in turn, requires that there is a chain of psychological connectedness over long periods of time. So, X considered at one time is psychologically continuous with Y, considered at another time, just in case there is a

⁸Mammal brains may exhibit something like this structure across hemispheres, albeit with significant biological interdependence between hemispheres. If our hemispheres could function as they do while being metabolically independent, then we would have a biological version of the case.

chain of psychological connectedness between X and Y. On this view, X is the same person as Y.

It's also natural to take psychological unity conditions to help determine the constitution of a sentient being at a time. While this thesis is rarely discussed, it's not implausible that part of what makes a mental state or body part mine and not yours is that it stands in psychological-functional relations to other parts of my mind. In this way, psychological relations determine the synchronic constitution of sentient beings.

The psychological view receives initial plausibility from the observation that normative relations like *being responsible for a past action* intuitively require psychological continuity between, e.g., the responsible agent and the actor that committed the action. However, there are good reasons to think that the psychological view doesn't exactly capture the ethically important individuation conditions, especially when it comes to determining interests.

The three cases mentioned above also pose problems for the psychological view. Consider again the first case: Remote Alex with Alex-monitor. In this case, Remote Alex is plausibly a psychologically unified entity. The states of Remote Alex are psychologically connected over time because successive states of Remote Alex directly share many psychological traits. The states are also psychologically unified at a time in an ordinary way. However, there is also a sense in which the mental states of Alex-monitor are also psychologically unified with the mental states of Remote Alex, both at a time and over time. That's because Alex-monitor has introspective access or is otherwise causally sensitive to the states of Remote Alex.

I've suggested that the Remote Alex and Alex-monitor intuitively involves two distinct moral units, supported by the observation that we don't seem able to compensate Remote Alex by providing benefits to Alex-monitor. Yet, it seems that there is some pressure for the

psychological theory to treat these processes as jointly composing a single moral unit because of the connections across Remote Alex and Alex-monitor.

It might be objected at this point that because the connections are asymmetrical, the psychological theory can distinguish Remote Alex from a larger entity that encompasses the processes of both Remote Alex and Alex-monitor. Indeed, Sider 2018 explores a diachronic case with a similar structure, where only one phase has access to the states of the other phase. In Sider's view, this case involves distinct individuals that partly overlap. Can we ground a similar result in the asymmetric connections between Remote Alex and Alex-monitor?

Against this possibility, there are reasons to think that synchronic asymmetries in connections do not ground a division of moral units. After all, a single mind could plausibly exhibit these same asymmetries if it were suitably modular. In the human mind, modules are relatively insulated systems of mental processes, where other mental states outside the module are sensitive to the outputs of that module. In the human mind, modules do not generally exhibit the stark asymmetry that we see in the Remote Alex and Alex-monitor case, but it's plausible that even if a module did exhibit an absolute asymmetry, it still might not generate a distinct moral unit. If that's right, then asymmetrical psychological connections at a time are not straightforwardly handled by the psychological theory. Rather, closer investigation of the particular organization and content of the psychological states would be required in order to capture the important difference across these cases.

To see an additional obstacle for the psychological theory, let's consider the second and third cases together. The second case involves one instance of Remote Alex run across multiple distinct servers or machines. As I envision the case, it involves some of the Alex processes being run on one server, where the output of that process is broadcast to another

instance. Each instance communicates outputs of its own processes to other parts, which in turn take those outputs as inputs to subsequent processes, and so on. (Many artificial intelligence algorithms are executed this way currently because they are too large to be processed quickly on a single server.) Still, despite the spatial separation of processes, it seems plausible that with the right sort of connections, the disjointed processes could together constitute a moral unit—even if those processes can’t ground a single field of conscious experience.⁹

The third case involves multiple instances of Remote Alex that are run synchronously, where these instances use and update an informationally rich shared representation of the community that they assist. These distinct instances share many psychological connections because of their shared access to the representation of external states. We can suppose that each instance has a huge set of beliefs based on this representation. So, the mental states across instances exhibit high similarity and causal interdependence. Even so, despite the shared information, it doesn’t seem automatic that we have a single moral unit in this case. Rather, it seems that the details could be filled in such that each instance is a distinct moral unit.

These two cases together pose a problem for the psychological theory because it’s difficult to see how this theory could give the desired verdicts in both cases. If the psychological theory entails that connections across servers ground unity, such as in the second case, then it should say the same thing about the third case. However, what seems relevant is not simply the amount of psychological connectedness, but rather the kinds of mental states that are involved in those connections and what other connections there are. While massive informational connections don’t seem to contribute to moral unity, it’s plausible that scatter

⁹Compare the case to the experiences of a human whose hemispheres have been severed by the severing of the corpus callosum. It’s plausible both that this human has two fields of conscious experience, but also that they constitute a single moral unit. Cf. Nagel 1971 for a discussion of these kinds of cases. The case may also bear some similarities to the minds of octopi (Godfrey-Smith 2016).

connections of a different sort, perhaps including functional or practical relations, might suffice for moral unity.¹⁰

In that's right, then we need a way to distinguish the connections that count toward unity from those that don't. The psychological theory itself does not capture this distinction. Moreover, it's plausible that the 'format' of the connection isn't the key factor: whether the connection is psychological, informational, or functional does not on its own determine whether that connection counts toward moral unity. So, even if the psychological theorist modifies their account to select among psychological, informational, and functional connections, the account no longer seems to be properly psychological, in that whether the connection is psychological doesn't determine whether the connection counts toward moral unity.

The purpose of these three cases is to show that biological and psychological unity and continuity conditions fail to capture the individuation conditions of moral units, especially for artificial beings. The cases are practically relevant because artificial beings could easily be remotely subserved, run on virtual machines, subserved across multiple machines, or duplicated. Our existing theories, which have been tested primarily against imagined scenarios about human beings, are not designed to explain these strange cases. Even so, I believe these theories get some aspects right, such as the emphasis that biological theories place on embodiment and the emphasis that psychological theories place on certain mental relations. In the next section, I motivate an approach to thinking about moral individuals that improves on biological and psychological theories. Then, I will use this approach to show how to individuate moral units of all kinds.

¹⁰Perhaps there would be unity if one machine feels concern for the other machine and is disposed to inherit desires or intentions from it. The view I ultimately pursue emphasizes connections of this sort of practical import as crucial for moral unity.

4.4 THE SENTIENCE-FIRST APPROACH

I will now motivate a new approach to individuating moral units. In general terms, we want a theory that carves out morally special objects. While biological and psychological unity and continuity conditions may individuate objects of some type, I have suggested that we can't successfully individuate moral units merely in terms of those conditions. Neither notion of unity can be the basis of the general theory of moral individuation because they give the wrong verdicts, or no verdicts at all, in the cases discussed. So, while biological and psychological conditions can be morally important, those conditions matter morally because they correlate with other conditions that ground moral value. Following a prominent tradition, I assume that sentience is the ground of this kind of moral value, i.e., the significance of morally considerable interests.¹¹ So, I propose that facts about sentience, including valenced mental states or dispositions to have them, are the right sort of facts by which to individuate moral units. The strategy is to first look at patterns of sentience and then to carve out moral units according to the distinctive clusters that emerge from those patterns. This is the sentience-first approach.

We can already see initial motivation for the approach. We want a theory for individuating moral units. It's common to individuate objects according to unifying connections, such as causal relations. Further, it's plausible that relations of sentience generate morally significant value. I am simply taking these together.

¹¹The view that sentience is the ground of moral status is not uncontroversial. Other candidate grounds include: consciousness, rationality, self-awareness, or being a human. I think an approach that began from consciousness and focused on morally important unity would align with the sentience-first approach (because the morally important conscious states are the valenced ones). The approach I pursue also allots a special role to self-awareness, though it doesn't require self-awareness. Rationality and humanity are, I think, too restrictive and so are ultimately unhelpful in this context.

Here is further motivation for this approach. Sentience is the capacity to experience good or bad feelings, such as pleasure or pain. These are conscious states with positive or negative *valence*. It's highly plausible that, if sentience is morally significant, then there is a close connection between the valence properties of a felt state and the way that it's morally significant. We've already seen an initial application of this idea: in our imagined case of Alex, it's quite plausible that we know that it is pro tanto bad for Alex to be in negatively valenced conscious states and pro tanto good for Alex to be in positively valenced conscious states.¹² That's a claim about Alex's morally significant interests, and it flows straightforwardly from the way that Alex is sentient.

If the parallels between sentience and moral significance can be extrapolated, then we should be able to learn a lot about the content and weight of interests grounded in some system by examining sentience properties of that system. We can learn about the moral significance of a pattern of sentience by learning about the dispositional profile that the pattern exhibits. That which elicits bad feelings is bad, and so on. Importantly, as I argue in chapter 3, this reliance on sentience does not imply that only pleasure and pain matter morally, despite the fact that those are the most obvious sentience-involving properties of a sentient system. Because sentience-involving dispositions matter in general, we can countenance the moral significance of mental states that have valenced felt components, such as values, desires, regrets, wishes, anxiety, and so on. Supposing that these mental states are the plausible basis of the interests of an artificial or biological creature, the individuation conditions of moral

¹²Does this rule out masochism? That is, does this imply that pain can never feel good? No, in at least three ways. First, it could be that painful states can also be pleasurable. Second, it could be that pain states are sometimes accompanied by distinct states that are pleasurable. Third, it could be that pain states are sometimes positively rather than negatively valenced. If the third possibility is true in Alex's case, then at least some cases of pain are not pro tanto bad for Alex.

units are plausibly reflected in them.

I have argued that a morally considerable being's interests depend on its individuation conditions, whether that is synchronic constitution or diachronic identity. From another angle, it seems plausible that interests track patterns of sentience. The sentience-first approach simply completes the triangle: individuation tracks sentience, too.

At this point, the approach can take a stronger or weaker form. The weaker form just says: for whatever reason, those three things covary. If the weaker form is true, then the sentience-first approach of individuating moral units should be interpreted in *epistemic* terms: it's just a strategy for figuring out what the moral units are. In contrast, the stronger form says that patterns of sentience *metaphysically* determine both individuation conditions and interests. I prefer this view, partly because I cannot see what else would determine the individuation conditions of artificial sentience and partly because I find it to be explanatorily powerful. However, in what remains, I will be neutral on whether we give the approach an epistemic or a metaphysical gloss.

In the rest of this section, I will explain how patterns of sentience determine interests without presupposing identity. To that end, I distinguish three types of sentience, show how each type generates morally significant interests, and then show how patterns of sentience might determine the individuation conditions of moral units. (Compare this discussion to sections 1.3 and 3.3.)

A valenced, conscious mental event—a manifestation of sentience—comes in one of three increasingly sophisticated types. At the lowest level, we have *mere sentience*, the capacity for valenced experiences such as of undirected pain or pleasure. If we could make a petri dish containing a network of neurons instantiating pain, but where that network had no capacity

for (other) contentful mental states, then that petri dish would be a merely sentient system. Plausibly, the moral significance of this system is that we have reason to minimize its pain states, perhaps even by destroying it.¹³

The next two levels are more important, as they involve mental states that are plausibly necessary for a being to have interests in continuing to exist, in bodily states, or in events that involve it at other times. At the second level, there is what I'll call *directed sentience*. This includes mental states like urges, desires, and emotions, including hunger, fear, anger, or sadness. Such mental states involve positive and negative valence. When an urge is quenched, that's usually accompanied by positive affect. When a desire is frustrated, that is usually accompanied by negative affect. Beyond valence, instances of directed sentience either have cognitive content or are functionally sensitive to other conditions. That is how those states get to have satisfaction conditions (in the cases where they do). The general characterization of directed sentience is that it includes any state or disposition with valenced affect that is directed at or otherwise organized around other things beyond that mental state, such as other mental states, body parts, or the objects of desire. These mental states are richer than pain states and ground additional interests in the beings that have them.

I've supposed that artificial sentient beings will at least have directed sentience. They will feel certain ways about or toward certain things. With these mental capacities on the table, we can explain the more sophisticated interests that artificial sentient systems might have. As just one example, consider a desire. A desire involves dispositions to feel satisfaction or frustration depending on its satisfaction and plausibly generates an interest for that desire to be satisfied.¹⁴ The desire is thereby morally significant.

¹³Koplin, Carter, and Savulescu 2021 discuss the moral significance of brain organoids.

¹⁴In Chapter 3, I develop a non-hedonistic view couched in terms of sentience-involving dispositions on

While directed sentience is important, it is of limited use for individuating moral units. Attitudes such as desire on their own can easily be directed at conditions clearly not belonging to the moral unit whose desire it is. After all, I can desire things for you, and we don't think that makes it the case that we are of the same moral unit.¹⁵ To find the appropriate joints for determining bodily interests and interests over time, I suspect we must introduce a more sophisticated version of directed sentience that I call *self-concern*. Self-concern involves valenced conscious states with self-directed representational content. A being with this capacity can refer to itself in thought, can form preferences about its future, and can feel certain ways about its past.¹⁶ Moreover, such a being can have opinions in how it qua itself is treated, such as about what physical interventions count as bodily contact, and by extension, which interventions are seen as violations of bodily security. It's plausible that some such attitudes and dispositions could give us highly informative clues about the actual bodily boundaries of a sentient being.

On the picture I'm sketching, both directed sentience and self-concern ground interests in systematic ways, and it's more complicated than in the case of mere sentience. With directed sentience or self-concern, for a being to have an interest in some event E, it must have some attitude or feeling that is sensitive to whether or not E transpires. That amounts to the subject being disposed to have some subjective response if they observe E or otherwise learn that E occurred. Note that some constituents of events play regular roles in the cognitive-

which the satisfaction of a desire counts in the interest of sentient beings even if they are unaware. While that non-hedonistic construal of the significance of desires is popular, the account to follow does not depend on it.

¹⁵Perhaps the presence of such a desire goes *some* way toward unifying us for moral purposes. If I desire benefits for you, then benefits to you count as benefits to me. That's a one-off instance of the same sort of moral relation that moral units characteristically bear to themselves. The point here is just that such desires cannot suffice on their own to individuate units.

¹⁶As Singer 1993 (p. 97) says, "Continued existence cannot be in the interests of a being who never has had the concept of a continuing self."

affective dispositions of the being or system in question. The bodily and self-involving interests of the being are no different. I will try to show how to ‘read off’ the embodiment of a being from these dispositions.

One way to do this is to find the objects or conditions that play the same roles in the artificial pattern of sentience as the human body plays in a typical human person’s pattern of sentience.¹⁷ I’m assuming that the body elicits sentience in a special way in human persons, and that this special directed pattern of sentience grounds our bodily interests. If we can find the objects or conditions that figure in the same way (at a suitably abstract level) in other sentient systems, then we should expect those systems to have similar body- or self-involving interests.

How does the body figure in a special way in our pattern of sentience? There are a few prominent aspects to how our body figures in our mental life. First, we have sensory and perceptual acquaintance with the conditions of our body, including through the various aspects of tactile sensation (taste, touch, heat transfer, various kinds of pain) and proprioception. Second, we have (non-sensory) cognitive attitudes about our body parts, such as a desire not to be touched in certain ways or beliefs that the hand in that photo is mine. Third, there are also agential relations, including motivations, intentions, and the execution of actions, which are also accompanied by characteristic feelings and evaluations. Typically, these aspects are intermixed in any episode of mental life. So, let us say that, if an object plays the aforementioned sensory, cognitive, and agential roles in a system of sentience-involving dispositions, then that object is a *target* of that system. This notion of a target includes mental states and body parts because such states and parts will figure in each of these roles to some

¹⁷I intend for my notion of ‘body part’ or ‘condition of the body’ to include mental states too.

extent. If we take the collection of such parts, given a system of sentience, then we may have a kind of provisional ‘whole’ body.

One characteristic feature of targets of system of sentience is that the system tends to exhibit various *preservation* dispositions toward its targets, which includes: retraction responses, such as if I recoil in pain from a hot stove; avoidance responses, such as if I am inclined not to look over the edge of the skyscraper; caretaking responses, such as if I want to tend my wounds; and so on. It’s plausible that these dispositions would be central even to artificial sentience.

The sentience-involving dispositions that generate the target roles are a plausible basis for the interests that track individuation conditions. For example, the way a body part figures in my pattern of sentience is closely connected, and perhaps even partly constitutive of, my interests regarding that body part. Plausibly, the fact that I would feel pain if I stubbed my toe is part of what grounds my interest in not stubbing my toe. The fact that I would be afraid of dying and the fact that dying would prevent the enjoyment of future events plausibly explain why I have an interest in not having my life threatened. These are interests that derive from my bodily constitution and from my diachronic identity conditions (the conditions that determine what it takes for me to survive), and it seems that we have a way of reading off these interests from the targets of my patterns of sentience. The hypothesis of the sentience-first approach is that we can read off similar interests, and individuation conditions, of perhaps any other sentient being in an analogous way.

For example, we might suppose that in the case of Remote Alex, the remote processes instantiate a feeling of pain when I press on a sensor in my home. Then, we should say that this sensitivity grounds Remote Alex’s bodily interest in me not pressing that sensor. The

sensor plays a role in the sentience-involving dispositions of Remote Alex in the same way that my toe plays a role in my sentience-involving dispositions. Because we think my sentience-involving dispositions ground my interest in not having my toe subjected to pain, we should infer by analogy that something similar is true in the case of Remote Alex. In other words, for moral purposes, we should treat that sensor as a body part of Remote Alex, where Remote Alex then has an interest in the condition of that body part. The sentience-first approach therefore settles what standard theories of identity were silent about: the question of whether Remote Alex has a bodily interest in the condition of sensors that are remotely located from its mental processes.

In the next section, with this notion of a target and the foregoing sketch of sentience-involving dispositions, I will show in more detail how the sentience-first approach can tell us how to individuate moral units.

4.5 APPLYING THE APPROACH TO CASES

The sentience-first approach to individuating moral units suggests the following strategy. First, given a set of mental dispositions to have valenced experiences, take the targets of those dispositions. Then, infer the individuation conditions of the moral units from the structure and content of these dispositions and their targets. In this section, I explain a version of this strategy in more detail by applying it to both artificial and biological cases.

As mentioned earlier, a special worry about artificial sentience is that the programs and processes that generate artificial sentience may be easily duplicable because the same programs and processes can be run multiple times across distinct machines. Like the third case discussed in section 4.3, let's consider a case where there are several instances of Remote Alex

that are nearly just alike qualitatively, and where they have a shared representation of the environment. On some versions of the psychological theory, these machines will be psychologically unified because (nearly) all of their mental states are similar and because they exhibit some degree of causal interactivity. On other versions of the psychological theory, such as those that require overlap in the realizers of mental states, the machines may be considered as belonging to distinct units. How can the sentience-first approach settle the case?

The key factor is that each instance of Remote Alex has token sentience-involving dispositions that are sensitive to the states of that token instance of Remote Alex (along with the sensors located in faraway homes) but not sensitive to the token states of other Remote Alex instances. As a result, it's generally true that providing experiential benefits to one will not produce corresponding benefits in another, nor is there any sense in which other instances will have a felt sensitivity to those benefits. In other words, the broader functional profile of felt sensitivity of each instance does not extend to include the states of other instances.

In contrast, consider again the second case, which described one instance of Remote Alex being run across multiple machines. The sentience-first approach says: in order for this distributed program to constitute a moral unit, there must be felt sensitivities across the disjoint machines. So, one machine must be disposed to feel or care about events that occur in the other machines. If there is no such sentience-involving sensitivity, then the case (for moral purposes) resembles a litter of puppies more than a person. Crucially, there need not be sensory acquaintance or conscious, introspective access across separate parts; rather, cognitive-affective relations such as self-concern suffice.

The remaining case is harder. This case is about Remote Alex joined with Alex-monitor. The trouble here is that Alex-monitor does have sentience-involving dispositions that are sen-

sitive to the states of Remote Alex, and so there is some sense in which the embodiment of Alex-monitor subsumes that of Remote Alex. In fact, the case I envision is not very different from a case in which there is simply one moral unit instantiated in these processes. What makes the difference?

Remote Alex is mentally sophisticated, not unlike a person. I have stipulated that the instance of Remote Alex in this case is functionally just like other instances. Because Remote Alex is quite sophisticated, it has the capacity for self-concern. That is, Remote Alex has a concept of self and characteristically directs feelings and other felt attitudes at itself by deploying that concept. In fact, in a system with this level of mental sophistication, it's plausible that the array of sentience-involving dispositions would be yoked together and organized by the functions of the self-concept. As a result, the sentience-involving dispositions instantiated in this whole system (Remote Alex together with Alex-monitor) contain a natural joint: there is a package of functionally interrelated dispositions that are sensitive to the states of Remote Alex but not sensitive to the states of Alex-monitor, and there is a distinct package of dispositions that are organized around the states of both Alex-monitor and Remote Alex. Remote Alex's dispositions are yoked together by its self-concept, but because that instance is not sensitive to the states of Alex-monitor, it does not incorporate those states as targets.

If Alex-monitor is similarly sophisticated, then it will have its own self-concern governed by its own self-concept. Now, the structure of Alex-monitor's dispositions is unusual: it is sensitive to the states of Remote Alex in much the same way that you are sensitive to the states of your foot. That is, it has some kind of internal or introspective access. However, the felt experiences of Alex-monitor will be distinct from the experiences of Remote Alex because those experiences have different properties than Remote Alex's experiences. (For ex-

ample, Alex-monitor might experience pain, experience Remote Alex's thought that "I am in pain," and then instantiate an *additional* thought that it is in pain.) If Alex-monitor is so sophisticated in this way, then it is plausibly a distinct moral unit. If that's right, then inducing pain in this Remote Alex system is twice as bad as inducing the same amount of pain in an unmonitored instance.

I'll now briefly discuss two interesting cases of human moral units to show that the sentience-first approach gives plausible verdicts in those cases too. Then I will get a little more precise about how the general form of the sentience-first strategy.

Consider a case of a dicephalic parapagus twins—conjoined twins that share most of their body but have two heads. Because these twins have distinct experiences and distinct senses of self, they should be classified as two distinct persons. The foregoing approach gives this verdict because, even though there is significant overlap in the targets of some of their mental states (e.g., each twin is sensitive to the condition of their shared feet), the targets of their dispositions don't fully overlap. One twin will have slightly different concerns and sensitivities than the other, including concern about distinct mental states that are not shared, and so the view implies that they are two distinct moral units.

The second biological case is that of a split-brain human—a human who has had its corpus callosum cut, severing its hemispheres. It's plausible that in such a case, there are two separate fields of conscious experience, two distinct token concepts of self, and slightly different dispositions of sensory acquaintance. So, one hemisphere might sense only the right hand, while the other senses only the left. Still, I argue that the sentience-first approach rules these distinct fields of consciousness as belonging to the same moral unit—the same person. How?

In this case, it's crucial that the token self-concept located within each hemisphere is disposed to be sensitive to the conditions of the whole body. That is true despite the lack of sensory acquaintance with half of the body. It's true, despite that lack, because cognitive-affective relations (such as thinking about the unfelt hand as one's own in a stable way) overcome the lack of sensory acquaintance. And so, because sentience-involving dispositions can have targets by way of cognitive relations, the targets of the sentience-involving dispositions of either side ultimately overlap, even if they aren't targets in exactly the same ways across the hemispheres.

I contend that these are the correct verdicts, and that the spirit of the sentience-first approach delivers those verdicts. But it's helpful if we can say something more precise so that the approach can be generalized and assessed against other cases.

The rough strategy is to take a set of sentience-involving dispositions, locate the targets of each of those dispositions, and then begin lumping the targets together according to the functional profile of feelings that are involved. Some of the sentience-involving dispositions will have multiple targets, or a big aggregated target. Fear of an incoming projectile, coupled with background beliefs about the body, has each body part and the aggregate body as targets. At the extreme, deployments of the self-concept may have the entire persisting person as its target. In systems with these mental properties, the individuation conditions of the moral unit are given by the largest aggregated target, the whole target of the self-concept (or more rudimentary cognates).

A little more precisely, we can use the following recipe. Take a set of sentience-involving dispositions instantiated in some system in there is an unknown number of moral units. Some of these dispositions will be yoked together in the following way: there will be a disposi-

tion that *inherits* the targets of other dispositions. Then, the target of the former disposition is at least as large as the target of the latter dispositions. Consider the strictly ‘largest’ targets and the associated dispositions. In sophisticated systems, these dispositions will be instances of self-concern. The largest distinct targets (together with the dispositions) constitute distinct moral units.¹⁸

Now, the recipe could be finished here with “and there are no other units in the system.” However, that would seem to preclude the verdict that there are two units in the case of Remote Alex and Alex-monitor because Alex-monitor’s self-concept would seemingly inherit all of Remote Alex as a target, thereby subsuming Remote Alex because the former is strictly larger than the latter. Admittedly, this case may evoke middling or uncertain intuitions from the reader, or it may seem to depend on unspecified details. But supposing that there is some plausible scenario in which Remote-Alex and Alex-monitor are distinct moral units, how could the recipe deliver that verdict?

As I see it, we must also recognize that *exclusiveness* in Remote Alex’s sentence-involving dispositions to the states of Alex-monitor is relevant. The exclusion makes it the case that Remote Alex is distinct from Alex-monitor. The fact that Remote Alex’s self-concept excludes the states of Alex-monitor makes it the case that Remote Alex is a distinct moral unit. (Indeed, this exclusion is part of the explanation as to why benefits to Alex-monitor would not count as benefits to Remote Alex.) So, it’s not enough that the ‘larger’ dispositional profile inherits the targets of the ‘smaller’ dispositional profile. In order for the smaller unit to be subsumed as part of the larger moral unit (and not a distinct moral unit), the smaller dispositional profile must not be explicitly exclusive of any of the targets of the larger dispo-

¹⁸Notice that the ultimate individuating step applies at the level of targets, not dispositions. Typically, the sentence-involving dispositions are included because they are also targets.

sitional profile. For that distinction to work, we need a notion of functional exclusion. It's clear enough how that works with the self-concept because there can be attitudes about parts being not one's own body parts. How that works in more rudimentary systems may present difficult edge cases, and such cases may require a more specific version of the foregoing recipe.

4.6 LIMITATIONS

Even though I've applied the approach to several cases, it faces important limitations. In particular, it's quite likely that some artificial sentient machines won't exhibit the stable sentience-involving structures that I have stipulated in the cases. It could be, for example, that sensitivities change over time or form a big, complex web of connections where there is no single individual constituted by the web. So, it might be that artificial sentience cannot be individuated into discrete moral units at all.¹⁹ With such systems, moral risks and moral uncertainty may be inevitable.

On a more optimistic note, it may be possible to develop a moral outlook that can helpfully and accurately guide our judgments about how to interact with such systems. Such a moral outlook may look more like Parfit-inspired connectedness-based ethics than the more ordinary identity-based ethics. I'm hopeful that something in the foregoing account will transfer to those stranger cases, but I leave that for future work.

¹⁹Thanks to Eric Schwitzgebel for pressing me on the importance of this possibility.

5

Conclusion

In this dissertation, I have developed an explanation of the metaphysical and normative distinctiveness of persons and other moral beings. Ultimately, the theory holds that the moral significance of creatures like us is grounded in patterns of reflexively directed affect, such as self-concern. In that way, what is morally valuable about us is grounded in what is intimately subjectively valuable to us.

Instead of saying more about what I have done, let me say something about questions that deserve more attention. While I believe that the theory succeeds in cases where it straightforwardly applies, what is most interesting to me about the foregoing account is how it might be helpfully applied to divergent cases.

So, what should we say about persons whose patterns of self-concern change drastically over time, contrary to the more optimistic assumptions of 1.4? If someone comes to reject a past self, or to fear anticipated transformation as a kind of mental corruption, what reasons do they—and we—then have regarding their interests over time? Would it be imprudent for a prescient and still-intact Phineas Gage to save for retirement? I suspect that cases like this may generate genuine prudential and moral conflicts. In that way, the moral collisions produced by fellow-travelers such as personites may reappear in a more substantial yet sparser form. I take these to be real possibilities worth exploring—possibilities not ruled out by the global buttressing to our moral outlook afforded by the self-concern account.

Or, what should we say about cases where conflicts in the self-concept are especially obstinate, despite the promise of unity discussed in 2.6? One possibility is that the self in that case is deeply fragmented. Another possibility is that there are two or more overlapping yet genuinely distinct selves, showing that pluralism may only be dispatched sometimes. What sets these cases apart? Arguably, the presence of a unifying disposition of the self-concept, when not fully exercised, is what makes a being fragmented despite the perseverance of the inconsistent attitudes. In cases where even the disposition is not present, then there are multiple distinct selves (or moral units). There may also be middle cases where the unifying disposition is neither determinately present nor determinately absent. Ambiguities such as these could generate interesting normative puzzles.

I argue in Chapter 3 that some formerly conscious humans are morally significant. But what of the others—those whose patterns of concern did not treat permanent unconsciousness as survival? Perhaps there is no moral reason to keep such humans on life-support. But what do we say if the family regards them as still there? Or what if the individual themselves did not have a settled opinion on the matter? Ambiguities like these can be distressing, and my account cuts against consolation.

As I mention at the close of Chapter 4, it's plausible that artificial sentience may not be divisible into discrete moral units. Indeed, the days of moral individuals may be numbered! Instead, the Earth could become covered in undifferentiated mycelium of morally significant mentality. What is good or evil when the many are one? Would this be a horrible future, or a beautiful one? Maybe the sentientist-first approach will tell.

Spanning these questions is an issue that I have left to one side: the sociality of the self. It seems to me that if the account articulated in Chapter 2 (and discussed a little more in the Appendix) is correct, then facts about my embodiment partly depend on facts external to me. In particular, I suspect that the values of people around me play a role in determining my nature. The causal version of this claim—that we are conditioned by and inculcated into a way of life—is hard to doubt. But a more radical constitutive claim is enabled on the foregoing account. I've said that my identity is determined by my pattern of concern. Suppose that I value being my partner's partner, where that requires fulfilling some of their values. Then, I value fulfilling their values. And so, I value being that way *de re*. If so, then facts about what my partner values may play a role in determining what I am. In that way, the self-concern explanation may show how people are socially constituted. The view may also reveal a new moral significance to that sociality.



Notes on a Metaphysics of Persons

The goal of this appendix is to show how the resources articulated in chapter 1 could be developed into a theory of personal identity. I also suggest that this metaphysics of persons could support moral status essentialism and perhaps even person essentialism.

A.1 OVERVIEW

In chapter 1, I argue that BEING THE OBJECT OF A PATTERN OF SELF-CONCERN is the distinctive property that grounds our ethical significance. However, there are two issues that deserve more discussion. First, it's not obvious how such a property relates to personal identity, and so it is not yet clear that persons are the bearers of such a property. Indeed, I leave open in chapter 1 that something shorter-lived than a person, such as a 'self', is the privileged object of ethical significance. Second, it's open whether our ethical significance—especially our moral status—is contingent or necessary. I take the view that our moral status is essential, and hence necessary, to be desirable, yet there are recent arguments against this idea (Olson and Witt 2020, Harman 2021).¹

In this appendix, I work toward an account of personal identity on which the ground of our ethical significance is essential to us. Because this feature is essential to us and grounds moral status, the view entails that our moral status is essential. I do not provide a sustained argument for the view, but I mention some features and advantages. I discuss some theoretical choice points without taking sides.

There are two overarching versions of the view. The first version holds that BEING THE OBJECT OF A PATTERN OF SELF-CONCERN is our fundamental essential kind, such that something with that essence could not exist without there being an instance of self-concern.² The second version is more inclusive, holding that BEING THE OBJECT OF A PATTERN OF REFLEXIVE AFFECT is our fundamental essential kind. Self-concern is one form of reflexive affect, and so the latter view is a sort of abstraction from the former. On that view, I could

¹For other work on this question, see Olson 1997, Rudder-Baker 2000, and DeGrazia 2005.

²The nearest predecessors to the view I develop are that of Johnston 1989 and Johnston 2010, especially chapter 4. See also Zimmerman 2013 for a translation of that view into a perdurantist ontology.

have existed as a late-term fetus or infant even if I would not later develop the capacity for self-concern (which requires the capacity for self-representation). I'm inclined to believe that the inclusive essential condition captures an essential feature of moral beings in general. For those who think that a deep moral distinction between persons and non-persons is written into the essences of things, the restrictive view of our essence will prove more attractive. The restrictive view notwithstanding, I will often speak simply of reflexive affect, or a pattern of reflexive affect, because this is the more inclusive term. I will distinguish the views explicitly when some issue bears on them differently (mostly in sections A.3 and A.4).

On either version of the view, our essence is determinable—it can be satisfied in various determinate ways. The fact that our essence is determinable accommodates variable embodiment, persistence despite attitudinal change, individual differences in embodiment, and a novel way to explain moral significance in terms of essence. Our determinable essence has these upshots because it induces a distinctive way that a property can be had: not essentially, nor accidentally, but *quasi-essentially*. Much of this sketch revolves around this notion.

A.2 TOWARD AN ESSENTIALLY SELF-CONCERNED BEING

A.2.1 DETERMINABLE ESSENTIAL CONDITIONS

A determinable property is a property that can be instantiated in multiple determinate, mutually exclusive ways. Canonical examples include the colors. Something that is red (a determinable) must be red in a more determinate way, such as by being the shade of red that is Boston University Red.

Determinable properties are neutral about their determinates. If a determinable property is essential to an object, this neutrality enables two important features that are character-

istic of ordinary objects. First, the determinates themselves may not be essential to the object. The essence of a person might say something about having mental states without saying anything about any determinate mental states, or it may say something about being an organism without saying anything about the specific biological processes involved (so, it may be neutral on whether one must have a functioning thyroid). Second, this neutrality enables objects to be capable of variable embodiment. I am essentially embodied, but *this* fully determinate embodiment is not essential to me. I shall survive the loss of some of these molecules, neurons, or attitudes, as well as the gain of others. That is, my embodiment varies over time. Even so, I satisfy the same essential determinable (having some body or other) at every moment of my existence.³

Determinables often have graded determinacy. Colors are again a helpful example. First, there is the maximally indeterminate condition of BEING COLORED. There are near subtypes, such as BEING WARMLY COLORED, followed by intermediate subtypes such as the ordinary colors (e.g., BEING RED). At the other end of the spectrum, so to speak, there are the fully determinate shades, such as BEING BOSTON UNIVERSITY RED or BEING CRIMSON. We can think of these types as corresponding to progressively finer natural regions of the color solid, which depicts every determinate shade organized by the subjective similarity relations of hue, saturation, and lightness.

The maximally indeterminate determinable class that figures in the essence of moral beings is that of *reflexive affect*, and one subtype of this class is that of *self-concern*. If there

³A natural way to regiment talk of variable embodiment of concrete particular objects is to appeal to a kind of essential functional role of the object, where the role condition is neutral over realizers. So, the essence of the object is satisfied when there is some (perhaps other) object that realizes that role. The object that realizes the role may not be the same object over time, but the role-object could persist nevertheless. An extreme example is exemplified in the TV show *Altered Carbon*, where some people can get new human bodies. Role-functional essences are a special case of the picture of determinable essences that I am sketching.

is a way of representing the space of all determinate forms of reflexive affect in some high-dimensional space, I couldn't tell you what its dimensionality would be. The dimensionality of color is 3 because the determinate shades can vary along dimensions of hue, saturation, and lightness. Instances of reflexive affect can presumably vary in valence, intensity, duration, felt character, representational content, functional profile (which includes both elicitation conditions and resultant conditions), and more. These parameters may be multidimensional in turn, and so the space could be quite rich. Notice also that the dimensions of reflexive affect are finer-grained than phenomenal distinctions (because functional profiles are not always phenomenally distinguishable), perhaps unlike color.

A moral being is essentially the object of a pattern of such instances. But which such instances are included in the pattern? That simple question will have complicated and messy answer which I shall only partly provide here. Very roughly, the answer is that instances of reflexive affect (or self-concern) belong to the same pattern when they are integrated in the right way, but the details must wait until section A.2.3.

Supposing we can identify the subjective dispositions that are included in a pattern, what then makes something its object? In section 4.4, I discuss the 'target' of subjective states, which we may provisionally think of as a general term corresponding to the way that, e.g., body or mental parts, events, or states figure in the subjective life of a sentient being. Note that there may or may not be some abstract general feature had in common by all 'targets', though the theory will be better if it turns out that there is. One candidate shared feature is that targets tend to be endorsed, identified-with, cared-for, protected, or otherwise included by some felt sense of *ownership*.⁴ Let us use this characterization of targethood, with the

⁴de Vignemont 2020 argues that the sense of ownership consists in the affective awareness of the importance of the body for survival. Shoemaker 2016 discusses a notion of ownership as central to personal identity, but

caveat that it may not be a reductive condition of targethood. The idea, then, is that the object of the pattern is the aggregated target, where targets are aggregated when there is some mental state in the pattern that treats those targets as parts of a whole target. At the limit, the overarching ‘self-concept’ (or a more rudimentary analog) stitches it all together (as described in Chapter 2).

There are two important aspects to how the pattern determines its object. First, the object is determined holistically. The ‘wide’ functional profile of a mental disposition includes other mental states, and the target of a mental state can be inherited via such indirect functional connections. Moreover, sometimes there may be conflict between various one-off judgments (such as I explore in chapter 2), and some of the putative targets of such states should not be considered genuine targets. To determine whether some putative target is really to be included as part of the object, a degree of coherence must be imposed on the overall pattern. (This is an ‘internal’ standard of propriety and not an independent standard, and so it does not undermine the response-dependence of the view.)

Second, the target of a particular mental state needn’t be related by *acquaintance* to that mental state. That is, something can be a target of a mental state even if no mental state discloses sensory, perceptual, or phenomenal aspects of that target. Rather, the target relation and the derived ‘object of’ relation can be grounded in purely cognitive relations, where the target is the intentional object of that cognitive state. This allowance is crucial in permitting that unremembered past stages (such as infants) and unconscious future stages (such as an anticipated being in a persistent vegetative state) can nevertheless be incorporated into the object of a pattern of self-concern (cf. section 3.5). This allowance also plays a role

notes that it’s various forms may not be a unified kind.

in permitting that artificial entities with multiple distinct fields of conscious experience can nevertheless belong to one and the same moral unit (cf. sections 4.4 and 4.5).

Now we are in a position to see how there gets to be a determinate embodiment. The body parts (including mental states, etc., but I will talk only of “body parts” for simplicity) that figure as the target of some part of the pattern get taken up in the aggregate object, which is itself the target of at least some deployments of the self-concept (or a more rudimentary analog). These body parts thereby indirectly satisfy the essential condition of the moral being, for which there must be some such object. The fully determinate embodiment, then, includes the complete intrinsic qualitative profile of that object, and so the embodiment also includes the pattern of mental states (whether self-concern or simpler forms of reflexive affect).

Many of the determinate features of an embodiment are not the specific targets of any affect-laden mental state. Rather, those determinate features may only be targets by virtue of satisfying a determinable condition, which itself is directly implicated by the mental state. For example, I am indifferent to whether my hand is composed of these molecules [holds up hand] or... [waiting a few seconds] these molecules [holds up the same hand again]. That is true despite the fact that I care about having that hand. Indeed, there are many properties or objects that are part of the ground of my essential conditions, yet for which I would shed no tear. That is how I get to have a variable embodiment on the essentialized self-concern view. To bring this point into focus, we must get quasi-essentiality on the table.

A.2.2 QUASI-ESSENTIAL CONDITIONS

The account exploits another property of determinables—one that I have not seen discussed elsewhere. Sometimes philosophers distinguish between essential and accidental features of

an object, where an object has its accidental features contingently and its essential features necessarily. Yet, determinable essential conditions induce a third category that I call *quasi-essentiality*. A condition φ is quasi-essential to x just in case x could have failed to be φ and where φ is part of the grounds of x 's existence. A condition φ is part of the grounds of the existence of an object if and only if it satisfies an essential condition of that object (such that the nearest possible duplicate-minus- φ of x does not meet the essential condition and hence is neither x nor a counterpart of x). So, the conditions that are quasi-essential to x are the conditions that are not essential to x but which satisfy the essence of x . This scheme is characteristic of determinable and disjunctive essential conditions.

To see the idea in action, consider a stop sign. Suppose a stop sign has three essential conditions: a stop sign is essentially red, octagonal, and displays "STOP". Then, for something to be a stop sign, it must be red. However, for something to be red, it must be a determinate shade of red, such as by being Boston University Red. The satisfaction of some such determinate is necessary for the satisfaction of the determinable. Yet, the stop sign is not essentially Boston University Red. It could have been, or could become, a different shade of red. So, what we have is a case where the object is contingently Boston University Red, but its being Boston University Red satisfies an essential (and hence necessary) condition for being that object. So, being Boston University Red is part of the grounds of the existence of the stop sign even if the stop sign could have failed to be Boston University Red. On my taxonomy, being Boston University Red is quasi-essential to the stop sign.

The distinction between accidental and essential properties is sometimes characterized like so: accidental properties are the properties that an object can survive losing, whereas that is not so for essential properties. This characterization also shows that quasi-essential

properties are a bona fide class along the same modal dimension. For, it's not unqualifiedly true that objects can survive the loss of their quasi-essential properties. You cannot merely subtract what makes the stop sign Boston University Red without destroying it. You must at the same time preserve its redness by making it red in a new way. On the other hand, it also not unqualifiedly true that the stop sign cannot survive the loss of its quasi-essential properties. For there is a way of replacing (perhaps gradually) the determinate red-making features of the sign with different red-making features. Having a property quasi-essentially, then, is a distinctive way of having a property.

In general, non-essential determinates of essential determinables are quasi-essential to the object. Yet, there is one further aspect to point out here, which will become important when I discuss the relationship between our quasi-essential properties and our moral significance. Unlike the notion of essentiality and accidentality, which are absolute notions, quasi-essentiality is often gradable. A natural way to regiment the grade of quasi-essentiality of a condition is by measuring the range of counterfactual adjustments to that condition that nevertheless preserve the essential condition. The gradability of quasi-essentiality derives from the gradability of determinables. If some paint is essentially colored but not essentially any particular color, then the quasi-essentiality of its being red is, in some sense, stronger (i.e., "more essential" or "less accidental") than the quasi-essentiality of its being Boston University Red. That difference is reflected in the fact that the range of ways of being colored by not being red is smaller than the number of ways of being colored by not being Boston University Red. That comparative fact is, in turn, due to the fact that a higher proportion of the space of colors is red than is Boston University Red.

The gradability of quasi-essentiality is relevant because I conjecture that the inherent

moral significance of some condition depends on how ‘strongly’ quasi-essential that condition is to a moral being.⁵ At the extremes, the essential conditions are as morally significant as the continued existence of the being, whereas the purely accidental conditions are not morally significant at all. In the middle, it’s sensible that the moral significance of some condition varies according to how essential it is, which is what quasi-essentiality is supposed to capture. To give some *prima facie* support to that idea, note that my hand is more important to me, and more ‘strongly’ quasi-essential to me, than the specific collection of molecules that compose it. The range of counterfactual scenarios in which I lack that hand (all else equal) is smaller than the range of scenarios in which I lack the same hand exactly as it is currently composed. To demonstrate the point differently, notice that the fact that I have a hand is more directly implicated by my essence than the fact that I have a hand with some specific molecular composition. The essential condition of being the object of a pattern of reflexive affect (or self-concern) must be satisfied by a determinate pattern of mental states and a determinate object of those mental states. My pattern of mental states rather directly implicates me having a hand as a part, but it only indirectly implicates me having a hand with this specific composition. The latter condition is involved only because it is a determinate of the former determinable. For the same reason, violations of the determinate condition are less morally significant than violations of the more determinable condition.

A.2.3 INTEGRATION

Earlier, I raised the issue of what it takes for some mental states to constitute a “pattern”. The short answer is that they must be functionally integrated in the right way. The longer answer

⁵The degree of quasi-essentiality is well-defined only if there is a privileged measure over counterfactual adjustments to the condition. More work is required to develop and defend this conjecture.

is not so neat. I will sketch enough of the picture to suggest that it can be filled in successfully.

I can think of two ways that integration might occur. First, let's say that mental state m_2 is included in the pattern of mental state m_1 whenever m_1 is disposed to inherit the targets of m_2 . Then, m_1 exhibits a sort of deference to the targets of m_2 . For example, suppose I think of myself by default as having two feet, and I also am currently worrying about retirement. The latter mental state will be disposed to incorporate the former, such as if I consider how much money to budget for my future footwear collection. That is, my projections about myself into the future will operate with background assumptions about myself that are inherited from other attitudes or mental states in my overall web of reflexive mental states. In other words, the attitude m_1 defers to attitude m_2 to fill in the details of what I take myself to be like. Then, any pattern that includes token attitude m_1 also includes m_2 .

The second mode of integration is when two sets of mental states have exactly the same token targets, even if they are not directly functionally related. The basic idea, and motivation, is this: a person could have two rather functionally isolated brain regions, such as if their brain hemispheres are totally severed. There could then be mental states, such as I-thoughts, that occur in both hemispheres. What I claim is that, if these mental states have the same targets, then they are included in the same pattern. Or, similarly, if a digital mind were subserved twice over but where each instance has all the same bodily or mental targets, then we should say that they are included in one and the same pattern—two minds of the same morally considerable being. Note that in this case it is permitted that each token mind needn't have their targets in the same way. One 'mind' may be acquainted with a limb from the inside, whereas the other merely cognizes that limb as its own. That is a way of having the same target, which is grounds for being of the same pattern.

A.2.4 MEREOLGY

What has been said so far gives us a rough idea of the embodiment of the moral being, but there are other features that are also determined by a pattern of reflexive affect. First, there is the mereological structure of the object, which I briefly address here. Then, there are temporal and modal properties, which I say something about in the next two sections respectively. (I also discuss temporal properties in 1.3 and 1.4.)

Targets may have independent mereological structure, but the pattern of reflexive affect contributes additional mereological structure. So, my left hand and my left forearm are each parts of my body, they are contiguous with and connected to each other, and they directly causally interact in such a way to compose a unified part (my lower left arm). Those mereological relations are more or less independent of any mental relations. Even so, aspects of my reflexive affect or self-concern may constitute additional connections between these parts, such that the way that they fit together and the way they figure in my whole embodiment partly depends on such mental states.

More starkly, *whether* some object is a part of a creature's body will depend on the pattern of reflexive affect. So, whether a limb is a part of a person's body depends on whether it is treated as a body part by their pattern of self-concern. That is true for prosthetic limbs as well. In that way, the view can deliver results about the mereological composition of entities that diverges from biological conditions or conditions of acquaintance from the inside.

At this point, it is also worth mentioning an advantage that falls out of this result. A central theme of this dissertation is that the property of persons I have highlighted (that of being self-concerned) and the correlative account of our essence enables especially good explanations of our moral significance. And again, at the outset of this appendix, I noted

that the view enables a novel way to explain moral significance in terms of essence. With quasi-essentiality and mereology on the table, I can now spell this point out in further detail.

Body parts are morally significant. They are protected by rights of bodily autonomy, and we tend to have a significant interest in their preservation. Even so, it doesn't seem that facts about humanoid organisms, metabolic dependence, or acquaintance from the inside adequately explain the moral significance of the body. For, some people have a condition known as "body integrity identity disorder" (BIID), where they desire the amputation of a limb and are distressed by its continued presence. The desire and distress characteristically associated with BIID are apparently caused by a mismatch between the individual's body image and their physical body (Bayne and Levy 2005). A body image is a mental representation of the shape and structure of one's body. It's plausible—even compelling—that in a morally relevant sense, the rejected body part of a person with BIID should not be regarded as a body part. Indeed, that is the reasoning of Bayne and Levy who argue that the medical amputation of such body parts is morally permissible. In their culminating support of this claim, Bayne and Levy write: "in an important sense, a limb that is not experienced as one's own is not in fact one's own" (ibid. p. 85).

There are two ideas, latent in the argument of Bayne and Levy, that are uniquely exemplified by the account I am developing here.⁶ The first idea is that one's bodily constitution is partly settled by facts about one's experience of candidate body parts. The second idea is that the moral significance of some candidate body part, including the significance of amputations thereof, depends on whether the candidate part is "in fact" a body part. In particular, Bayne and Levy argue that surgeons should be allowed to amputate limbs that are mentally

⁶Compare this discussion to a similar account of prosthetic embodiment of Aas 2021.

rejected due to BIID. In a similar spirit, the account I develop holds that one's embodiment is partly determined by one's pattern of self-concern, which is shaped in part on one's body image. Further, the account I develop holds that much of our moral significance is explained by that pattern, and so it delivers the verdict that the rejected limbs are not subject to moral protections in the way that true body parts are. Thus, my account provides a metaphysical justification for the conclusion of Bayne and Levy (2005).

I would also add that the explanation is especially satisfying because it appeals to people's essential constitution. On the determinable picture I have articulated, the degree of quasi-essentiality explains how morally significant that condition is. In particular, the stringency of the connection between a pattern of reflexive affect (or, in the case of persons, self-concern) on the one hand and the target body part on the other explains the significance of that body part. Explanations that bottom out in essential facts are often seen as especially satisfying because there is no further question about why that particular object figures in the explanation in the way that it does.

A.2.5 TEMPORALITY

Attitudes and other mental dispositions can have temporal import too, either by having temporal content or directedness. Basic examples include memories and anticipations, but any retrospective and prospective attitude can contribute to the temporality of reflexive affect. Attitudes such as beliefs that I was a certain age or that I was born from that person (indicating to my mother) help make it the case that I am something that existed at that time, and attitudes such as my plans for retirement or my commitment that irreversible comas constitute death help make it the case that I will exist after I retire but not if I become irreversibly

comatose.

Perdurantists are welcome to interpret such contributions in mereological terms, i.e., as making it the case that some future event is a temporal part of my existence. Endurantists must interpret the contribution differently, perhaps in terms of a profile of persistence conditions or historical properties.

A.2.6 MODALITY

Attitudes and other mental dispositions can have modal import, too. When you decide not to go to graduate school because you think it would be bad for your mental health, you are presupposing a certain possibility: that you could go to graduate school. That possibility is counterfactual, and your implicit endorsement of it as your possible condition thereby helps fill out your modal profile. More generally, the range of possibilities that you see as compatible with your possible existence are possibilities in which you could in fact exist, unlike those possibilities that you see as incompatible with your possible existence.

If I think I could survive passing through a blackhole, does that make it true? We'd better hope not. No: there are further constraints, such as that many of my concerns are directed at having a functioning body. A possibility must cohere with one's entire pattern of concern. The important caveat to reiterate here is that some particular attitude within the overall pattern of reflexive affect doesn't automatically chisel away the modal profile of its possessor. Rather, the contribution is settled holistically, in concert both with other attitudes and with the relevant facts of the matter. See also 2.6 for further discussion.

A.2.7 CONSTITUTION & CIRCULARITY

On this view, a moral being is constituted by some body parts that are interrelated in the right way (especially to mental dispositions of reflexive affect, self-concern, etc.). I have said repeatedly that a person is the object of a pattern of self-concern. This has been a white lie. The object—the embodiment—is a constituent of the essence of a person, and I'd like the account to be reductive, if possible. So, the constituents of the essential conditions must be metaphysically prior to the thing itself, and hence the person cannot simply be (in the sense of identity rather than mere predication) that object. Officially, the person is constituted by that embodiment on the condition that the embodiment is related in the right way to the pattern of self-concern.

This produces a sort of oddity in our metaphysical bookkeeping. No doubt, you are capable of thinking about yourself. That is, you are a constituent of the content of some of your thoughts. It's typical to hold that a thought in some sense depends on what it's about. But then, your first-personal thoughts are metaphysically dependent on you. However, we cannot maintain in general that your reflexive attitudes are dependent on you in this way because I have already committed to the claim that you are dependent on those attitudes. How can we avoid problematic circularity?

To avoid circularity, we must 'subtract' any dependence on the person that is present in their mental states. What remains may legitimately figure in the grounds of the existence of the person. Is there a well-motivated way to do this?

Yes. Note that many of the relevant attitudes are about our parts, and so those attitudes pose no threat of circularity. Even so, there is a question about how these parts get to be integrated as parts of the whole self or person without there being a mental state that has

the whole as part of the content. I think the right solution is to hold that there are some metaphysically prior mental states that are unifying by expression or functional directedness and not by way of content. So, a pattern of mental states treats parts as parts of a larger integrated whole without representing them as such. I think that is enough to solve the problem.

A.2.8 REFLEXIVITY

The account appeals to attitudes and other mental dispositions that I have called “reflexive.” Strictly speaking, a relation is reflexive just in case it relates a thing to itself. That is not how I have used the term. Similarly, you might think that self-concern—as a kind of reflexive relation—is a relation that, necessarily, is had only between a self and itself. Again, that is not how I have used the term.

The notion of reflexivity (and the derived notion of self-concern) in the foregoing account is more inclusive. As I use the notion, an attitude or mental disposition can count as reflexive by being directed at a suitably ‘nearby’ part, such as another mental state to which it is closely functionally connected or a body part to which it has some kind of internal access. On the other hand, some instances of self-concern are reflexive in the paradigm sense, at least when directed at the whole self or person.

What makes an attitude or disposition reflexive in the broader sense? A mental disposition can be reflexive in (at least) two ways. First, it can be reflexive in a basic way when the subjective response is a kind of sensory acquaintance with the internal states of the relatum. So, my feeling of pain when stubbing my toe counts as an instance of reflexive affect because of the subjective response is a kind of internal sensory acquaintance with the condition of the toe.

Second, a mental disposition is reflexive when the relatum of the subjective response is also a target of the broader web of mental dispositions. What makes my concern for my future pain a reflexive attitude is that the attitude is integrated with a number of other attitudes that have that anticipated future pain as a target. So, this second way of being reflexive depends on the broader web of mental states with which the mental state is functionally integrated. In general, reflexive states are those that hold between a part of a pattern of mental dispositions and a part of the aggregated target of that pattern.

A.3 DISSOLVING A PROBLEM OF DETERMINATION

I have occasionally worried about a putative problem for self-concern essentialism. The problem is that, because it takes a couple years for young humans to develop self-concern, then self-concern essentialism faces a dilemma: in order to preserve commonsensical verdicts about when persons come into and go out of existence, the view must hold that facts about when a person comes into existence are either retroactively determined or socially determined.

Retroactive determination is quite puzzling. On that version of the view, for any human baby, there is currently no fact of the matter about what entity that baby is, and yet, one day, it will be the case that this baby is (say) Newt. Beyond this head-scratching entailment, retroactive determination is incompatible with the thesis that we persist by enduring—an incompatibility I aim to avoid.

On the other hand, social determination is incompatible with the intrinsic determination of identity over time. Some theorists hold that whether y is the same person as x should depend only on the intrinsic state(s) of y (and x). In cases where $x=y$, this amounts to the persistence of x being secured by the intrinsic state of x over time. Indeed, this ‘only x - y rule’ or

‘only x rule’, as it is sometimes called, is a metaontological commitment about the persistence of concrete particulars. Because social determination of when a person came into existence is extrinsic to that being, it violates this rule. While some theories of personal identity violate that rule (e.g., Nozick 1981), that may be undesirable.

However, it turns out that both of these options can be avoided. The best route for avoiding retroactive or social determination depends on which version of the view we prefer—whether our fundamental essential kind is that of being the object of a pattern of reflexive affect or, instead, of self-concern.

A.3.1 REFLEXIVE AFFECT ESSENTIALISM

The first option holds that our fundamental essential kind (the kind that settles our absolute persistence conditions) is that of being the object of a pattern of reflexive affect. On that view, we come into existence as infants or late-term fetuses—long before and independently of whether we develop the first-personal representational capacities that are required for genuine self-concern. So, the view preserves commonsensical claims about our original existence without appealing to retroactive or social determination.

A nice feature of this option is that it commits to the same fundamental essential kind for all morally considerable beings. All God’s creatures, then, are fundamentally the same kind of thing. Other additional benefits are that the view is compatible with our endurance or perdurance, as well as a version of moral status essentialism.

It’s important to note that, even on this version of the view, persons and other moral beings can persist despite losing the capacity for reflexive affect or consciousness in general—as discussed in chapter 3. The moral being comes into existence when there is a suitably

organized pattern of reflexive affect, but it can take on an embodiment that transcends this condition.⁷

A.3.2 MIXED PERSISTENCE

The second option, in short: concede that infants are not identical to adults (i.e., not $x=y$) but that infants nevertheless are the same person as the adults. This will be palatable to ‘ex-durantists’ (that is, stage-theorists) and perdurantists. This option enables a humbler variety of retroactive or social determination that is not as controversial as the sort discussed at this beginning of this section.

On this version of the view, facts about personal identity will be either retroactively or socially determined. What makes it the case that some infant in the past is the same person as me will depend on attitudes that are not intrinsic to that infant at that time or any prior time. Rather, other people’s attitudes (on the social route) or attitudes I developed later (on the retroactive route) are what make me the same person as that infant. Even so, that kind of determination is not so rare or controversial. In general, perdurantists about personal identity such as Lewis (1976) subscribe to some such notion of retroactive determination. For example, on Lewis’s view, the number of people coinciding with you right now depends on future facts about fission. If retroactive determination of personal identity is unpalatable to you, you should adopt a social determination picture or be an endurantist and pursue the ‘reflexive affect’ version of the view.

⁷Whether this specific verdict is compatible with endurantism is subtle, turning on how past facts figure in the present constitution of formerly conscious beings. If past objects such as former selves are parts of the person as it continues to exist, then the view is not compatible with endurance. If the past figures in some other way, such as by being the truthmaker to a historical property of the person, then it may indeed be compatible with the endurance of the being.

Even so, here is an olive branch for the endurantist who may, for whatever reason, feel that the reflexive affect view does not work. There could be parts or aspects of the person that are identical over time and which thereby do persist by being one and the same thing located at multiple distinct times. The natural candidate is something like a psychologically-individuated ‘self’ or personality, which occupies a smaller spatiotemporal region than the whole (perduring) person. This view of mixed persistence is not unique to the current proposal (cf. Paul 2017). This component of the view may placate those who find endurantism about persons compelling.

A.4 PERSON ESSENTIALISM AND MORAL STATUS ESSENTIALISM

In this section, I show how the foregoing discussion about the metaphysics of persons paves the way for a version of moral status essentialism. I then suggest how that thesis might be adjusted to capture something closer to person essentialism—at least if we adopt the mixed persistence option of A.3.2.

Moral status essentialism about an entity is the claim that the entity has moral status as a necessary consequence of its essential features. As I understand the term, an entity has moral status just in case its interests are a source of agent-neutral moral reasons for the sake of that entity. Then, we say that its interests are morally considerable.

The view, or family of views, that I have sketched so far can support moral status essentialism. The basic idea is this: entities that are essentially the object of reflexive affect (or, in some cases, self-concern) thereby also have interests that are morally considerable. Whenever the essential conditions for such a being are satisfied, the being thereby exists, and its interests will be morally considerable.

The thesis of moral status essentialism must be argued for on two fronts, descriptive and normative. On the descriptive side, it must be shown that the posited condition is indeed plausibly an essential condition of the entities in question, which we might argue for by establishing that such a condition is extensionally plausible. On the normative side, the strategy is to argue from general moral principles to the claim that reflexive affect suffices for grounding morally considerable interests. The normative strategy is quite close to the project pursued in chapter 3, in which I argue that this sort of picture is a well-motivated extension of sentientism. I won't pursue that line of argument further here.

The descriptive side of the argument requires at minimum that the view does not deliver wildly implausible verdicts about the existence and persistence of any morally considerable creatures. It strikes me as plausible that humans, intelligent alien beings, animals, and so on, qua morally considerable beings, all come into existence when they instantiate reflexive affect. That claim implies that you did not exist as an early fetus but that you existed by infancy. Given that the view holds that the moral beings are those that are the objects of reflexive affect rather than the subjects of reflexive affect, the view is also compatible with the claim that you could continue to exist in a permanent unconscious condition such as a coma (if that condition is included under the horizon of your self-concern). I take it that those verdicts are plausible, even if they are not exactly the verdicts you would have otherwise endorsed.

Even though the view supports a version of moral status essentialism, it's important to note that the view does not entail that morally considerable beings have their determinate interests essentially. Thus, the moral significance of a being is contingent and may vary over time. Given this, one might find the supported version of moral status essentialism to be

rather weak. It's worth exploring whether a stronger version of this view is also supported by the foregoing metaphysics. In particular, it's worth exploring whether the view supports person essentialism, where that is understood as the view that any being with full moral status (the sort of moral status characteristic of beings with moral personhood, which consists in rights and other special protections) has that status because of its essential profile. (Note that I favor talk of 'inherent moral significance' over 'moral status'. See section 3.6.)

On its face, person essentialism seems quite implausible. After all, moral personhood seems to be possessed by beings with psychological personhood, i.e., the capacity for self-conscious reflection (Locke 1975, Olson and Witt 2020). Because you existed as an infant and hence at a time when you did not have the capacity for self-conscious reflection, and because you could have failed to develop the capacity for self-conscious reflection, then person essentialism must be false of you—and every other known person. What is the hope for securing this view?

The synchronic version of person essentialism is indeed implausible, as argued by Olson and Witt (2020). However, there is a nearby diachronic variant available on perdurantist theories of persons, according to which full moral status is earned for a being even at times when it doesn't satisfy the Lockean psychological criteria of personhood. Roughly, the view is this: persons are essentially the (four-dimensional) objects of patterns of self-concern, and any being that meets this condition thereby has (full) moral status. The condition can be met even at times when the capacity for self-conscious reflection is lacked, such as by infants, because the pattern of self-concern as it is instantiated later can include the infant. The view then implies that, e.g., something cannot be me unless it is the object of a pattern of self-concern, and so whenever I exist I satisfy the condition required for having full moral status.

The view may seem paradoxical, but it can be clarified by stating it in standard perdurantist terms. Call the temporal part of my life when I was an infant, considered just as it was in intrinsic categorical terms, my ‘infant stage’. It’s contingent whether my infant stage grows into something with the capacity for self-concern. It’s also contingent whether my infant stage is (a part of) me. What isn’t contingent is that *I* am the object of self-concern, and by closure over entailment, that I am a person in the moral sense.⁸ Though the view may seem strange, the strangeness derives from the perdurantist picture. That may be good reason to pursue the endurantist alternative stated in terms of reflexive affect, in which case person essentialism must be relinquished.

⁸It’s here that perdurance-enabled person essentialism gives different verdicts than the ever-conscious view of Harman 2021.

Listing of figures

2.1	Figure 2.1	49
2.2	Figure 2.2	54

References

- [1] Aas, S. (2021). Prosthetic embodiment. *Synthese*, 198, 6509–6532.
- [2] Adams, R. M. (1989). Should ethics be more impersonal? A critical notice of Derek Parfit, *Reasons and Persons*. *The Philosophical Review*, 98(4), 439–484.
- [3] Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, 63(4), 596–612.
- [4] Baker, L. R. (2000). *Persons and Bodies: A Constitution View*. Cambridge University Press.
- [5] Bartels, D. M. & Rips, L. J. (2010). Psychological connectedness and intertemporal choice. *J. Exp. Psychol. Gen.*, 139-49.
- [6] Bartels, D. M. & Urminsky, O. (2011). On intertemporal selfishness: How the perceived instability of identity underlies impatient consumption. *Journal of Consumer Research*, 38(1), 182–198.
- [7] Bayne, T. & Levy, N. (2005). Amputees by choice: Body integrity identity disorder and the ethics of amputation. *Journal of Applied Philosophy*, 22(1), 75–86.
- [8] Bennett, K. (2004). Spatio-temporal coincidence and the grounding problem. *Philosophical Studies*, 118(3), 339–371.
- [9] Berniūnas, R. & Dranseika, V. (2016). Folk concepts of person and identity: A response to nichols and bruno. *Philosophical Psychology*, 29(1), 96–122.
- [10] Blok, S., Newman, G., & Rips, L. J. (2005). Individuals and their concepts. In R. L. G. W. Ahn, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin* (pp. 127–149). Washington, DC: Am. Psys. Association.

- [11] Bostrom, N. & Shulman, C. (ms). Propositions concerning digital minds and society. unpublished.
- [12] Bostrom, N. & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (pp. 316–334). Cambridge University Press.
- [13] Braddon-Mitchell, D. & Miller, K. (2004). How to be a conventional person. *The Monist*, 87(4), 456–474.
- [14] Broome, J. (2004). *Weighing Lives*. Oxford University Press.
- [15] Bruckner, D. W. (2013). Present desire satisfaction and past well-being. *Australasian Journal of Philosophy*, 91(1), 15–29.
- [16] Butler, J. (2005). *Giving an Account of Oneself*. New York: Fordham University Press.
- [17] Chen, S. Y., Urminsky, O., & Bartels, D. M. (2016). Beliefs about the causal structure of the self-concept determine which changes disrupt personal identity. *Psychological Science*, 27, 1398–1406.
- [18] Das, N. & Paul, L. A. (2020). Transformative choice and the non-identity problem. In A. Sauchelli (Ed.), *Derek Parfit's Reasons and Persons: An Introduction and Critical Inquiry* (1st ed.). Routledge.
- [19] de Vignemont, F. (2020). What phenomenal contrast for bodily ownership? *Journal of the American Philosophical Association*, 6(1), 117–137.
- [20] DeGrazia, D. (2005). *Human Identity and Bioethics*. Cambridge University Press.
- [21] Degrazia, D. (2007). The harm of death, time-relative interests, and abortion. *Philosophical Forum*, 38(1), 57–80.
- [22] DeGrazia, D. (2021). An interest-based model of moral status. In S. Clarke, H. Zohny, & J. Savulescu (Eds.), *Rethinking Moral Status*. Oxford University Press.
- [23] Dietz, A. (2020). Are my temporal parts agents? *Philosophy and Phenomenological Research*, 100(2), 362–379.
- [24] Dorsey, D. (2013). Desire-satisfaction and welfare as temporal. *Ethical Theory and Moral Practice*, 16(1), 151–171.

- [25] Dorsey, D. (2017). Why should welfare fit?? *Philosophical Quarterly*, 67(269), 685–24.
- [26] Dover, D. (2022). The conversational self. *Mind*, 131(521), 193–230.
- [27] Eklund, M. (2020). The existence of personites. *Phil. Studies*, 177, 2051–2071.
- [28] Faden, R. R., Beauchamp, T. L., Mathews, D. J. H., & Regenberg, A. (2021). 159C10 Toward a Theory of Moral Status Inclusive of Nonhuman Animals: Pig Brains in a Vat, Cows versus Chickens, and Human–Nonhuman Chimeras. In *Rethinking Moral Status*. Oxford University Press.
- [29] Fairchild, M. (2019). The barest flutter of the smallest leaf: Understanding material plenitude. *Philosophical Review*, 128(2), 143–178.
- [30] Fairchild, M. & Hawthorne, J. (2018). Against conservatism in metaphysics. *Royal Institute of Philosophy Supplement*, 82, 45–75.
- [31] Finlay, M. & Starmans, C. (2022). Not the same same: Distinguishing between similarity and identity in judgments of change. *Cognition*, 218(104953).
- [32] Fodor, J. A. (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford University Press.
- [33] Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, 40, 351–401.
- [34] Godfrey-Smith, P. (2016). *Other minds: The octopus, the sea, and the deep origins of consciousness*. Farrar, Straus and Giroux.
- [35] Greaves, H. (2019). Against “the badness of death”. In Gamlund & Solberg (Eds.), *Saving People from the Harm of Death*. Oxford, USA: Oxford University Press.
- [36] Guin, U. K. L. (1974). *The Dispossessed*. Harper Voyager.
- [37] Hare, R. M. (1981). *Moral Thinking: Its Levels, Method, and Point*. Oxford: Oxford University Press.
- [38] Harman, E. (1999). Creation ethics: The moral status of early fetuses and the ethics of abortion. *Philosophy and Public Affairs*, 28(4), 310–324.
- [39] Harman, E. (2021). The Ever Conscious View and the Contingency of Moral Status. In *Rethinking Moral Status*. Oxford University Press.

- [40] Hawley, K. (2001). *How Things Persist*. Oxford: Oxford University Press.
- [41] Heathwood, C. (2005). The problem of defective desires. *Australasian Journal of Philosophy*, 83(4), 487–504.
- [42] Heathwood, C. (2016). Desire-fulfillment theory. In G. Fletcher (Ed.), *The Routledge Handbook of the Philosophy of Well-Being* (pp. 135–147). Routledge.
- [43] Heathwood, C. (2019). Which desires are relevant to well-being? *Noûs*, 53(3), 664–688.
- [44] Heiphetz, L., Strohminger, N., & Young, L. L. (2017). The role of moral beliefs, memories, and preferences in representations of identity. *Cognitive science*, 41(3), 744–767.
- [45] Heller, M. (1990). *The Ontology of Physical Objects: Four-Dimensional Hunks of Matter*. Cambridge University Press.
- [46] Hershfield, H., John, E. M., & Reiff, J. S. (2018). Using vividness interventions to improve financial decision making. *Policy Insights from the Behavioral and Brain Sciences*, 5(2), 209–215.
- [47] Hershfield, H. E. (2011). Future self-continuity: how conceptions of the future self transform intertemporal choice. *Annals of the New York Academy of Sciences*, 1235, 30–43.
- [48] Hershfield, H. E., Garton, M. T., Ballard, K., Samanez-Larkin, G. R., & Knutson, B. (2009). Don't stop thinking about tomorrow: Individual differences in future self-continuity account for saving. *Judgment and dec. making*, 4(4), 280–286.
- [49] Husserl, E. (1960). *Cartesian Meditations: An Introduction to Phenomenology*. Translated by Dorion Cairns. Kluwer Academic.
- [50] Jaworska, A. (2007). Caring and full moral standing. *Ethics*, 117(3), 460–497.
- [51] Jaworska, A. & Tannenbaum, J. (2014). Person-rearing relationships as a key to higher moral status. *Ethics*, 124(2), 242–271.
- [52] Johnston, M. (1989). Relativism and the self. In M. Krausz (Ed.), *Relativism: Interpretation and Confrontation* (pp. 441–472). Notre Dame University Press.
- [53] Johnston, M. (2010). *Surviving Death*. Princeton, NJ: Princeton University Press.

- [54] Johnston, M. (2016). Personites, maximality, and ontological trash. *Philosophical Perspectives*, 30(1), 198–228.
- [55] Johnston, M. (2017). The personite problem: Should practical reason be tabled? *Noûs*, 50(4), 617–644.
- [56] Johnston, M. (2019). Is hope for another life rational? In P. Draper (Ed.), *Current Controversies in Philosophy of Religion*, volume 1 (pp. 47–68). Routledge, 1 edition.
- [57] Johnston, M. (2021). The subject and its apparatus: Are they ontological trash? *Philosophical Studies*, 178(8), 2731–2744.
- [58] Joshi, P. & Fast, N. (2013). Power and reduced temporal discounting. *Psychological science*, 24(4), 432–438.
- [59] Kaiserman, A. (2019). Stage theory and the personite problem. *Analysis*, 79(2), 215–222.
- [60] Kamm, F. M. (2005). Moral status and personal identity: Clones, embryos, and future generations. *Social Philosophy and Policy*, 22(2), 283–307.
- [61] Knobe, J. (2022). Personal identity and dual character concepts. In K. Tobia (Ed.), *Experimental Philosophy of Identity and the Self*. Bloomsbury.
- [62] Koplin, J., Carter, O., & Savulescu, J. (2021). Moral status of brain organoids. In S. Clarke, H. Zohny, & J. Savulescu (Eds.), *Rethinking Moral Status*. Oxford University Press.
- [63] Korman, D. Z. (2020). *Ordinary objects*. Stanford Encyclopedia of Philosophy.
- [64] Korsgaard, C. M. (1989). Personal identity and the unity of agency: A kantian response to parfit. *Philosophy and Public Affairs*, 18(2), 103–31.
- [65] Kovacs, D. M. (2016). Self-made people. *Mind*, 125(500), 1071–1099.
- [66] Kovacs, D. M. (2020). Diachronic self-making. *Australasian Journal of Philosophy*, 98(2), 349–362.
- [67] Leckie, A. (2013). *Ancillary Justice*. Orbit, New York, 1st edition.
- [68] Leslie, S.-J. (2011). Essence, plenitude, and paradox. *Philosophical Perspectives*, 25(1), 277–296.

- [69] Lewis, D. (1976). Survival and identity. In A. O. Rorty (Ed.), *The Identities of Persons* (pp. 17–40). University of California Press.
- [70] Lewis, D. (1986). *On the Plurality of Worlds*. Wiley-Blackwell.
- [71] Liao, S. M. (2007). Time-relative interests and abortion. *Journal of Moral Philosophy*, 4(2), 242–256.
- [72] Liao, S. M. (2020). The moral status and rights of artificial intelligence. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence*. New York.
- [73] Lin, E. (2017a). Against welfare subjectivism. *Noûs*, 51(2), 354–377.
- [74] Lin, E. (2017b). Asymmetrism about desire satisfactionism and time. In M. Timmons (Ed.), *Oxford Studies in Normative Ethics*, vol. 7 (pp. 161–183). Oxford, UK: Oxford University Press.
- [75] Locke, J. (1689/1975). An essay concerning human understanding. In P. H. Nidditch (Ed.), *The Clarendon Edition of the Works of John Locke*. Oxford: Oxford University Press.
- [76] Maley, C. (2023). Analog computation and representation. *British Journal for the Philosophy of Science*.
- [77] Martin, R. (1997). *Self-Concern: An Experiential Approach to What Matters in Survival*. Cambridge University Press.
- [78] McCoy, J., Paul, L. A., & Ullman, T. (2019). Modal prospection. In B. M. A. Goldman (Ed.), *Metaphysics and Cognitive Science* (pp. 235–247). Oxford: Oxford University Press.
- [79] McMahan, J. (2002). *The Ethics of Killing: Problems at the Margins of Life*. New York: Oxford University Press.
- [80] McMahan, J. (2019). Early death and later suffering. In Gamlund & Solberg (Eds.), *Saving People from the Harm of Death*. Oxford University Press.
- [81] Mott, C. (2018). Statutes of limitations and personal identity. In T. Lombrozo, J. Knobe, & S. Nichols (Eds.), *Oxford Studies in Experimental Philosophy* (pp. 243–269). Volume Two. New York, NY, USA: Oxford University Press.
- [82] Nagel, T. (1971). Brain bisection and the unity of consciousness. *Synthese*, 22(3-4), 396–413.

- [83] Nichols, S. & Bruno, M. (2010). Intuitions about personal identity: An empirical study. *Philosophical Psychology*, 23(3), 293–312.
- [84] Ninan, D. (2021). *Williams on the self and the future*. Analytic Philosophy.
- [85] Noonan, H. W. (2010). The thinking animal problem and personal pronoun revisionism. *Analysis*, 70(1), 93–98.
- [86] Nozick, R. (1974). *Anarchy, State, and Utopia*. New York: Basic Books.
- [87] Nozick, R. (1981). *Philosophical Explanations*. Harvard University Press.
- [88] O’Driscoll, K. & Leach, J. P. (1998). No longer gage: an iron bar through the head. early observations of personality change after injury to the prefrontal cortex. *BMJ (Clinical research edition)*, 317(7174), 1673–1674.
- [89] Olson, E. (1997). *The Human Animal: Personal Identity Without Psychology*. Oxford University Press.
- [90] Olson, E. (2010). Ethics and the generous ontology. *Theoretical Medicine and Bioethics*, 31(4), 259–270.
- [91] Olson, E. T. & Witt, K. (2020). Against Person Essentialism. *Mind*, 129(515), 715–735.
- [92] Parfit, D. (1971). Personal identity. *Philosophical Review*, 80, 3–27.
- [93] Parfit, D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.
- [94] Parfit, D. (2012). We are not human beings. *Philosophy*, 87(1), 5–28.
- [95] Paul, L. A. (2017). The subjectively enduring self. In I. Phillips (Ed.), *Routledge Handbook of the Philosophy of Temporal Experience*. Routledge.
- [96] Persson, I. (2021). Moral status and moral significance. In S. Clarke, H. Zohny, & J. Savulescu (Eds.), *Rethinking Moral Status*. Oxford University Press.
- [97] Quilty-Dunn, J. (2021). Polysemy and thought: Toward a generative theory of concepts. *Mind and Language*, 36(1), 158–185.
- [98] Rovane, C. (1997). *The Bounds of Agency: An Essay in Revisionary Metaphysics*. Princeton University Press.
- [99] Schechtman, M. (1996). *The Constitution of Selves*. Cornell University Press.

- [100] Schechtman, M. (2001). Empathic access: The missing ingredient in personal identity. *Philosophical Explorations*, 4(2), 95 – 111.
- [101] Schechtman, M. (2014). *Staying Alive: Personal Identity, Practical Concerns, and the Unity of a Life*. Oxford University Press.
- [102] Schwitzgebel, E. & Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, 39(1), 98–119.
- [103] Schwitzgebel, E. & Garza, M. (2020). Designing ai with rights, consciousness, self-respect, and freedom. In M. Liao (Ed.), *Ethics of Artificial Intelligence* (pp. 459–479). Oxford University Press.
- [104] Sebo, J. (2022). Op-ed: What should we do if a chatbot has thoughts and feelings? <https://www.latimes.com/opinion/story/2022-06-16/artificial-intelligence-morals-ethics-sentience-thinking>.
- [105] Shoemaker, D. (2007). Personal identity and practical concerns. *Mind*, 116(462), 317–357.
- [106] Shoemaker, D. (2016). The stony metaphysical heart of animalism. In S. B. . P. Snowdon (Ed.), *Animalism* (pp. 303–328). Oxford: Oxford University Press.
- [107] Shulman, C. & Bostrom, N. (2021). Sharing the world with digital minds. In S. Clarke, H. Zohny, & J. Savulescu (Eds.), *Rethinking Moral Status*. Oxford University Press.
- [108] Sider, T. (2001). *Four Dimensionalism: An Ontology of Persistence and Time*. Oxford University Press.
- [109] Sider, T. (2003). Maximality and microphysical supervenience. *Philosophy and Phenomenological Research*, 66(1), 139–149.
- [110] Sider, T. (2018). Asymmetric personal identity. *Journal of the American Philosophical Association*, 4(2), 127–146.
- [111] Singer, P. (1977). *Animal Liberation*. Avon Books.
- [112] Singer, P. (1993). *Practical Ethics, 2nd edition*. Cambridge University Press.
- [113] Sinnott-Armstrong, W. & Conitzer, V. (2021). How much moral status could artificial intelligence ever achieve? In S. Clarke, H. Zohny, & J. Savulescu (Eds.), *Rethinking Moral Status*. Oxford University Press.

- [114] Smith, M., Lewis, D., & Johnston, M. (1989). Dispositional theories of value. *Aristotelian Society Supplementary*, 63(1), 89–174.
- [115] Starmans, C. & Bloom, P. (2018). Nothing personal: What psychologists get wrong about identity. *Trends in cognitive sciences*, 22(7), 566–568.
- [116] Strawson, G. (2011). *Locke on Personal Identity: Consciousness and Concernment*. Princeton University Press.
- [117] Strohminger, N., Knobe, J., & Newman, G. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science*, 12(4), 551–560.
- [118] Strohminger, N. & Nichols, S. (2014). The essential moral self. *Cognition*, 131(1), 159–171.
- [119] Strohminger, N. & Nichols, S. (2015). Neurodegeneration and identity. *Psychological Science*, 26(9), 1469–1479.
- [120] Sumner, L. W. (1996). *Welfare, Happiness, and Ethics*. Oxford University Press.
- [121] Taylor, C. (1985). Self-interpreting animals. In *Philosophical Papers* (pp. 45–76). Cambridge University Press.
- [122] Thomasson, A. L. (2007). *Ordinary Objects*. Oxford University Press.
- [123] Tierney, H. (2020). The subscript view: A distinct view of distinct selves. In *Studies in Experimental Philosophy* (pp. 126–323). Oxford University Press.
- [124] Tierney, H., Howard, C., Kumar, V., Kvaran, T., & Nichols, S. (2014). How many of us are there? In J. Sytsma (Ed.), *Advances in Experimental Philosophy of Mind*. Continuum Press.
- [125] Tobia, K. P. (2015). Personal identity and the phineas gage effect. *Analysis*, 75(3), 396–405.
- [126] Tooley, M. (1972). Abortion and infanticide. *Philosophy and Public Affairs*, 2(1), 37–65.
- [127] van der Deijl, W. (2021). The sentience argument for experientialism about welfare. *Philosophical Studies*, 178(1), 187–208.
- [128] Walsh, E. (2020). Cognitive transformation, dementia, and the moral weight of advance directives. *American Journal of Bioethics*, 20(8), 54–64.

- [129] Williams, B. A. O. (1970). The self and the future. *The Philosophical Review*, 79(2), 161–80.
- [130] Wolf, S. (1986). Self-interest and interest in selves. *Ethics*, 96, 704–20.
- [131] Yablo, S. (1987). Identity, essence, and indiscernibility. *Journal of Philosophy*, 84(6), 293–314.
- [132] Zewe, A. (2022). New hardware offers faster computation for artificial intelligence, with much less energy. <https://news.mit.edu/2022/analog-deep-learning-ai-computing-0728>.
- [133] Zimmerman, D. (2013). Personal identity and the survival of death. In F. Feldman & B. Bradley (Eds.), *The Oxford Handbook of Philosophy* (pp. 97–154). of Death.