Tim Campbell[1]

# Advance Directives and the Someone Else Problem[2]

*Abstract: Advance directives (ADs) are commonly defended as a way for individuals to extend their autonomy into periods of incapacity, such as advanced dementia. Yet in some cases, the person who signs the directive and the later patient with severe dementia appear to be "different people", in a way that challenges the claim that the directive has any authority with regard to the patient's care. This is known as "the someone else problem." In this paper, I distinguish two different versions of the someone else problem and respond to each, arguing that the authority of ADs can be preserved, even in cases involving "different people" in either of the two senses considered. Moreover, I present a thought experiment—a case of "reversed dementia"—to show that the signatory's preferences continue to have normative weight, even when they conflict with those of her demented future self.*

[1] Institute for Futures Studies, Stockholm. Email: timothy.campbell@iffs.se

[2] Funding from the Swedish Research Council (1-VR-041) is gratefully acknowledged.

# 1. Introduction

At age 57, Anna is diagnosed with early-onset dementia. Knowing the likely progression of the disease, she signs an advance directive stating that if she later enters a state of severe cognitive decline—unable to recognize loved ones, communicate meaningfully, or make autonomous decisions—she does not want life-sustaining treatment in the event of a life-threatening illness. Her directive states clearly:

> "If I am no longer capable of recognizing my family or understanding my surroundings, and if I fall seriously ill, I do not want to be kept alive through artificial means such as a ventilator or feeding tube. Please allow me to die peacefully."

Years later, a person who, in many ways, appears to be Anna has advanced dementia. This person does not recognize Anna's children or husband, and is confused about where she is. She doesn't recall signing any directive. Yet, she does not seem to *mind* being in this state. She is cheerful, enjoys listening to music, and responds positively to caregivers. Later, she contracts a life-threatening case of pneumonia. Her caregivers face a decision whether to withhold life-sustaining treatment, as Anna's directive states. If they do, then their patient will die peacefully. Otherwise, she will continue to exist in her usual cheerful state.

Cases like this illustrate what has become known as "The someone else problem". It has been claimed that advance directives (ADs) lack moral authority when the person issuing the directive and the later dementia patient are, in some apparently ethically significant sense, 'different people'.[3] By 'lacks authority', I mean that the directive does not generate a special kind of moral reason *regarding the person who issued the directive*, to honour her explicit wishes. For practical purposes, the signatory of the directive is gone, and the patient who remains is "someone else".

In this paper, I distinguish two versions of the someone else problem, explaining how each could be thought to undermine the authority of advance directives. I then address both versions of the problem. I argue that if the person issuing the directive (the signatory) is continuous with the later dementia patient in one or more morally significant ways, then this can provide the relevant kind of reason to honour the directive, i.e., to regard it as authoritative.

However, the strength of the reason may depend on the degree, and type, of continuity, and should not be regarded as decisive, since other ethically significant reasons may outweigh this reason, or at the very least, complicate the situation. For example, in cases like Anna's Directive, the signatory may be incapable of meeting

[3] See, e.g., Buchannan (1989), Dresser

the standard of informed consent. This is because she may be unable to fully grasp *what it is like* to be in a state in which many of her preferences, such as her preference for recognizing loved ones, are no longer held. She may also be unable to grasp what it is like to be in a state of having new preferences that she does not currently have, such as a strong preference for simple pleasures. She may lack informed consent even if she is given as much relevant information as possible prior to signing the AD (Walsh, 2020).

My aim in this paper is not to answer such questions but rather to argue that an AD can have authority, i.e., it can provide a moral reason, regarding the signatory, to honour her expressed wishes, even when the dementia patient is rightfully regarded as being 'someone else' in either of the ways distinguished below.

In Section 2, I distinguish the two versions of the someone else problem, which I label 'the numerical identity version' and the 'different selves' versions, and I further distinguish two interpretations of the latter version of the problem. In Section 3, I give reasons for thinking that the numerical identity version of the problem does not pose a significant challenge for the authority of ADs. I consider some of the different morally significant ways in which the dementia patient is continuous with the signatory and argue that some these forms of continuity are sufficient to ground a special reason to honour the signatory's directive. In Section 4, I address the different selves version of the problem. I argue that on the first understanding of this problem, it can be treated in exactly the same way as the numerical identity version, and that on the second understanding, it fails to provide a convincing case against the authority of ADs because an individual's past preferences can ground moral reasons to act against her (different) present preferences. I illustrate this with a hypothetical case of "reversed dementia". I conclude in Section 5.

## 2. The Someone Else Problem: Two Versions

The first version of the someone else problem understands the relationship between the signatory and the dementia patient as *numerical non-identity*; the signatory and the dementia patient are numerically distinct individuals. The descriptions 'the signatory' and 'the dementia patient' do not refer to one and the same individual at different times; instead, they refer to numerically distinct individuals.

The second version of the problem construes the relationship between the signatory at t1 and the dementia patient at t2 as one of different selves, where the relevant notion of 'self' is logically compatible with the signatory and the dementia patient being numerically identical. I discuss each of these versions of the problem in turn.

## 2.1 The numerical identity version

Consider an individual x, existing at a certain time, t1, and an individual y, existing at some later time, t2. For x-at-t1 and y-at-t2 to be *numerically identical* is for x-at-t1 and y-at-t2 to be one and the same individual. Numerical identity is different from *qualitative identity*, which refers to things being exactly the same in terms of the kinds of properties they instantiate. Derek Parfit illustrates the difference between numerical and qualitative identity with the example of two indistinguishable billiard balls: even if they are qualitatively identical (same size, color, mass), they are not numerically identical — they are still two distinct objects (Parfit, 1984, Section 76). Likewise, a person at one time and another person at a later time might be psychologically similar without being numerically identical.

As DeGrazia emphasizes, numerical identity is a strict, all-or-nothing relation: "If person A at t1 is numerically identical to person B at t2, then A and B are one and the same being" (DeGrazia, 2005, p. 27).[4]

On the numerical identity version of the someone else problem, the dementia patient at t2 is psychologically so disconnected from the signatory at t1 that the former and the latter are numerically non-identical. This assumes a *psychological criterion* of identity over time. A standard version of the psychological criterion is given by Parfit (1984). It involves two concepts—*psychological connectedness* and *psychological continuity*. Parfit defines *psychological connectedness* as the holding of particular direct psychological connections such as a memory of a past experience, or an acting out of an earlier-formed intention. He defines *strong psychological connectedness*, somewhat arbitrarily, as a number of direct psychological connections that is "at least half the number that hold, over every day, in the lives of nearly every actual person" (205). Finally, he defines *psychological continuity* as "overlapping chains of strong connectedness" (205). Individual X at time t1 can be psychologically continuous with individual y at (earlier or later) time t2 even if there are few, or no, direct psychological connections between x at t1 and y at t2.

The psychological criterion, as a criterion of numerical identity over time, says that for any x that is a person at time t1 and any y that exists at some (earlier or later) time t2, x-at-t1 and y-at-t2 are numerically identical *if and only if* (1) x-at-t1 and y-at-t2 are psychologically continuous (there are overlapping chains of strong connectedness between them), (2) this psychological continuity has the right cause, and (3) it has not taken a 'branching' form.[5] Parfit's version of the psychological criterion includes a

[4] Here, I shall leave aside the question whether it can ever be indeterminate whether numerical identity holds.

[5] This criterion differs in two ways from Parfit's (see 1984, 206). First, what he calls "The Psychological Criterion" explicitly includes a necessary and sufficient condition for when X at t1 and Y at t2 are psychologically continuous. The condition is: if and only if there are overlapping chains of strong connectedness. This condition is presupposed, but not explicitly stated in my criterion. Second, Parfit's criterion concerns when a person X at t1

further claim that numerical identity 'just consists' in the holding of facts such as (1)—(3). The psychological criterion is not merely a criterion of identity, it *is* the fact of identity.

Parfit is agnostic as to what 'the right cause' is. It could be the normal cause, a special type of cause invoked in laws of psychology, or even *any* cause whatsoever. Psychological continuity takes a 'branching' form when there is a least one time from t1 to t2 at which more than one individual is psychologically continuous (with the right cause) with x at t1 or y at t2. This includes instances of so-called fission of persons, which I discuss in Section 3.

According to the psychological criterion, if the signatory at t1 and the dementia patient at t2 are not psychologically continuous, then they are not numerically identical. This would be the case if there were some point in the progression of dementia such that for any time t prior to that point, and any later time t*, the person at t and the person at t* are not strongly psychologically connected; the chain of strong connectedness that is necessary for identity would then be broken.

According to the numerical identity version of the someone else problem, if the signatory at t1 and the dementia patient at t2 are numerically non-identical, it is difficult to see how the AD signed by the signatory at t1 carries any authority—how it provides any moral reason, specifically a moral reason *regarding the signatory and the expression of their wishes*, to do what the AD says. As Allen Buchanan puts it, "...it is not clear why the earlier person's wishes should have authority over the later one" (Buchanan 1988, p. 267). Buchanan's concern is that non-identity undermines the moral authority of the directive because it treats one individual's preferences as binding on another, a move that requires special justification which appears to be absent in the case of the signatory and the dementia patient.

One way of responding to this problem is to reject psychological criteria of identity. For example, David DeGrazia (1999, 2005) addresses the problem by adopting a version of *animalism*, the view that each of us is essentially and most fundamentally a human organism whose identity over time consists in biological rather than psychological continuity. On this view, 'the signatory' and the 'dementia patient' refer to the same human organism. There is a single individual that signs the AD at t1, and then later, at t2, suffers from dementia, though this individual's psychological traits (e.g. memory, values, personality) have changed significantly from t1 to t2 (DeGrazia, 2005, pp. 24–28).

---

and a person Y at t2 are one and the same person. There is a possible reading of 'same person' that is different from numerical identity. Following Olson (1997, pp. 22-23), my criterion is explicitly given in terms of numerical identity and does not rule out the possibility that a person at one time could be numerically identical to something that isn't a person at some later time.

On the basis of animalism, DeGrazia argues that the authority of advance directives is unchallenged: the directive applies to the same individual, despite psychological discontinuity (2005, pp. 41–45).

## 2.2. The different selves version

There is another version of the someone else problem which cannot be addressed by appealing to animalism, or to any non-psychological view of identity.

According to this 'different selves' version of the someone else problem, the signatory and the later patient with dementia may be one and the same individual, but the two stages of that individual's life, the one during which the directive is signed, and the one during which the dementia patient receives life-sustaining aid, involve distinct *selves*.

Das and Paul (2022) give an account of a "self" as a temporal part of an individual, specifically a temporal part of an individual that has a stable (unchanging) configuration of certain mental properties—core preferences, beliefs, and character traits—and that is maximal in the sense of being the longest-lived temporal part of the individual with that specific configuration of mental properties.[6] It is possible, on this account of a self, for an individual to have multiple selves. For instance, if we assume that the terms 'the signatory' and 'the dementia patient' refer to temporal parts of a human organism, then these terms could refer to different selves. Perhaps 'the signatory' refers to a temporal part of the organism with a set S of distinctive core preferences, beliefs, and other character traits, that remain stable over the period of its existence, while 'the dementia patient' refers to a later temporal part of the same organism with a very different set S* of mental properties. Even if S and S* overlap to some extent, there could be mental properties that are members of S but not of S*, and vice versa. For instance, perhaps the signatory has a very strong core preference for recognizing particular loved ones' faces, and does not care at all for birdwatching, while the dementia patient is completely indifferent regarding the recognition of those particular faces and has a core preference for birdwatching.

One need not accept a temporal parts ontology to claim that an individual can have (or be) different selves at different times. For instance, some animalists, such as

[6] Specifically, a self is a temporal part of a person with the following features:

1. Prudential Status: the self is a unit whose well-being can be an object of prudential concern, both for itself and for other selves of the same person.
2. Constancy: during the relevant time interval, the person's identity-conferring mental states (core beliefs, values, character traits) remain stable.
3. Maximality: the self is the longest such temporally unified stretch — it isn't just a momentary state but a maximal period during which the identity-conferring states remain unchanged.

Olson (2007) and van Inwagen (1990) reject a temporal parts ontology. But animalists of their stripe could still claim that a self just is a human organism *with* a specific configuration of preferences, beliefs, etc. A human organism may thus become a different self just as it may take on a different career; it might be a firefighter at one time and then become a philosopher at a later time. Similarly, it might be one self, and then later become a different self after changing its core preferences.

Although the signatory and the dementia patient may be numerically identical, the signatory at t1 may be a different self from the dementia patient at t2. One way of understanding the different selves problem is this: If the individual's mental profile at t1 is sufficiently different from her mental profile at t2, then the case may be *relevantly like* one involving different (numerically distinct) individuals. The relevant units of moral concern are here assumed to be selves rather than individuals. The challenge does not come from the metaphysics of identity, but from an intra-personal moral conflict that is similar in relevant respects to most interpersonal moral conflicts.

Another way of understanding the different selves problem is that individuals (not selves) are the units of moral concern, but when one becomes a different self, the preferences of her past self no longer provide moral reasons for satisfying those preferences, especially when this would involve acting against her current preferences.

There are cases which seem to support the claim that a person's past preferences do not provide a moral reason regarding that person to take actions that would frustrate her current preferences after she undergoes radical psychological change. For instance, consider the actual case, reported by Markowitsch and Staniloiu (2025), of a man (call him Richard), who suffered amnesia after a car accident that occurred while traveling with his family. A few days after the accident Richard could no longer remember his past. He did not recognize close relatives—his sons felt like strangers—and when one son tried to speak with him as "Dad," Richard insisted he use Richard's first name instead. His marriage unraveled. Richard treated his wife as if the shared history between them did not exist, and he ultimately divorced her. He then left the city where they had lived. His family tried to "bring him back" by confronting him with people and places from his past, but to no avail.

While situations such as this may be tragic, it seems there is no moral reason *regarding Richard* (as opposed to his family members) to force Richard back into his former family life, even if we assume that he had a very strong past preference to continue in that life.

This suggests that, for the same kind of reason, an advance directive does not retain its authority in cases of conflicting preferences between past and present (or future) selves.

# 3. Parfit's Fission Example and Numerical Identity

In this section, I argue that the numerical identity version of the someone else problem does not undermine the claim that the AD signed by the signatory lacks authority regarding how the dementia patient is treated.

There is independent reason to believe that the mere fact that x-at-t1 is numerically distinct from y-at-t2 is not necessary for x-at-t1 to have some authority regarding what happens to y-at-t2. This reason is based on Parfit's famous argument for the claim that 'identity is not what matters' in survival.[7]

Parfit's argument appeals to hypothetical cases of perfectly symmetrical "fission". In such cases, an individual undergoes a procedure whereby she appears to "split into" two different individuals. The fission case that Parfit develops is a version of a case due to David Wiggins (1967, 53-54). Since I am assuming animalism, I shall present a version of the Wiggins/Parfit case that is compatible with both biological theories (animalism) and psychological theories of personal identity.

Consider Ava, a human organism, normal in almost every biological respect: anatomy, metabolism, brain-based psychology, etc., but with one biological peculiarity: "amoeba-like" regeneration. If Ava's body is divided into sufficiently large viable parts, each can regenerate into a complete human organism, rebuilding missing organs and tissue to restore full functioning (including both biological and psychological continuity), in roughly the way a flatworm can regenerate, but at the scale and sophistication required for a human.

Suppose Ava's body has a latent "bilateral redundancy": she is functionally organized into two perfectly matched halves, left and right, each containing a structurally complete "half-brain" that is individually capable of supporting consciousness and memory.

Now consider *The Regeneration Case*: Ava is injured in a freak accident. A heavy object falls on her, cleanly dividing her vertically along the midline and instantly destroying her right half, leaving the left half intact. Because of Ava's special biology, the left half quickly regenerates, resulting in a complete human organism. When the process finishes, there is one living person who is biologically and psychologically just like Ava prior to the accident.

Intuitively, this is an extreme case of survival through injury and repair. Ava loses a huge portion of her body, but she survives and regenerates.

Now consider *The Fission Case*. This is like the Regeneration Case but with the following difference. A freak accident divides Ava cleanly into left and right halves, but neither half is destroyed. Each half regenerates into a complete human organism,

[7] Parfit 1984, pp. 245–280.

and each of those organisms, call them Righty and Lefty, is psychologically and biologically continuous with Ava prior the accident. Ava's division is perfectly symmetrical: there is no non-arbitrary fact we can point to in order to support the judgment that Ava is Righty but not Lefty, or vice versa. It would therefore be completely arbitrary to identify Ava only with Righty, or only with Lefty. Moreover, Ava is not numerically identical with *Righty and with Lefty*, since that would imply, by the transitivity of numerical identity, that Righty and Lefty are identical, which is false.

In The Fission Case we have two individuals, each standing in the same survival-like relation to the original Ava that the single regenerated individual stood in to Ava in The Regeneration Case.

The Parfitian argument for the claim that identity is not "what matters in survival" is roughly as follows. First, in The Regeneration Case, Ava survives the accident; what matters in survival is preserved in that case. In The Fission Case, although Ava does not literally survive (she is not numerically identical to anyone who exists after the accident), each of the two regenerated successors is continuous (e.g., psychologically and biologically) with Ava just as the single successor in The Regeneration Case is. There is no additional loss of psychological continuity, no added disruption, no extra gap—just two continuers instead of one. If what grounds the "survival" judgment in The Regeneration Case is the preservation of those continuity/connectedness relations, then, since those very same relations are present, twice over, in The Fission Case, it seems what matters is preserved in The Fission Case. How can having an extra continuer make it the case that what matters is not preserved? As Parfit rhetorically asks, "How can a double success be a failure?"

If Righty and Lefty each have what matters regarding Ava, then what matters cannot be numerical identity. More carefully, numerical identity cannot be a necessary condition for what matters. Instead, what matters (and what is necessary for what matters) must be some other relation, or relations.

We can now connect the Parfitian argument with the question whether identity is necessary for the authority of an AD. Consider a variation of the Regeneration and Fission Cases that involves dementia. Suppose that in Regeneration-Dementia, Ava signs an AD before the accident that destroys her right half. Later, after her regeneration, she becomes the victim of severe dementia and depends on life-sustaining aid provided by her caretakers. In Fission-Dementia, Ava signs an AD before her perfectly symmetrical division, and then one of her continuers (say, Lefty) is later the victim of severe dementia and becomes dependent on life-sustaining aid.

If Buchanan and others are correct that the authority of an AD depends crucially on identity, then even if Ava's AD carries authority in Regeneration-Dementia, Ava's AD carries no authority whatsoever in Fission-Dementia. In other words, there is no

special Ava-regarding reason to honour the AD that was signed by Ava. But intuitively, this is not the case. Ava's AD has at least some special reason-giving force; it cannot be completely undercut by the fact that Ava has Righty as an additional continuer.[8]

The intrinsic relations between Ava-at-t1 and Righty-at-t2 in Fission-Dementia are exactly like those between Ava-at-t1 and Ava-at-t2 in Regeneration Dementia. If Ava's AD in the second case carries authority, it is hard to believe that Ava's AD in the first case carries no authority whatsoever.

One might try to resist this conclusion by focusing on the literal language used in advance directives. For instance, many directives say things like "I do not wish to live in a state where I cannot recognize my loved ones," or "If I become unable to speak, walk, or feed myself, I do not want my life prolonged." If the signer and the future dementia patient are numerically distinct individuals, does this not render such first-person language misleading or even meaningless?

But to treat these linguistic conventions as decisive would be to miss the point. The arguments above suggest that when an individual at time t1 is continuous with a certain individual at time t2 in all the ways that would suffice for identity except for the presence of a second continuer at t2, then what matters is preserved, including what matters for the authority of an AD signed by the individual at t1. Emphasizing the use of "I" in the directive, and rejecting its authority because of a lack of numerical identity, would be an overly legalistic and superficial response. The moral weight of a directive should rest not on the semantics of its pronouns, but on the preservation of what matters.

Even if we accept the claim that identity is not necessary for the authority of an AD, we might think that 'what matters' as regards authority is not preserved in cases of severe dementia. It might be claimed, for instance, that what matters is psychological connectedness and continuity, and that in many cases of severe dementia the psychological connections between the signatory and the dementia patient are too few to preserve the authority of an AD.

I make several points in response.

First, psychological connectedness is a matter of degree, and even if the signatory and the dementia patient are not *strongly* psychologically connected in Parfit's sense, they may still be psychologically connected to some degree.[9] If psychological connectedness matters to whether an AD provides a moral reason *regarding the signatory* to do

[8] To avoid potential complications concerning what opinions Righty might have about what Lefty's caretakers ought to do regarding Lefty's treatment, we can assume that Righty dies in an unfortunate accident prior to the time when Lefty develops dementia, and is not around to opine about what should be done with Lefty. I assume this does not affect the claim that Ava is not identical to Righty.

[9] C.f. DeMarco, J. P., & LiPuma, S. H. (2016). For empirical evidence of the preservation of some psychological

as she desires, then insofar as connectedness is a matter of degree, the strength of that reason should also be a matter of degree, which may vary directly with the degree of connectedness. Hence, even if the signatory and the dementia patient are only weakly psychologically connected, this may only support the claim that the moral reason is weaker than in cases of strong connectedness.

Second, the signatory and the dementia patient may be continuous in other ways that matter. For instance, there may be physical, functional, and organizational continuity of the brain and nervous system sufficient for the continuity of the signatory's capacity for consciousness. As McMahan (2002), Unger (1990) and others have argued, this seems sufficient to ground *egoistic concern*; the signatory at t1 may care in an egoistic way about what happens to the dementia patient at t2, even assuming a loss of numerical identity.

Third, even biological or physical continuity may be sufficient to ground a moral reason to honour the signatory's AD. For instance, Govind Persad (2019) argues persuasively that one may retain ownership rights over one's body, rights that constrain, at least to some degree, what others may do to it even after one has ceased to exist, and even if one's body comes under the possession of someone else. Continuity of the body *itself* may not even be necessary for grounding the moral reason that is associated with whatever rights one normally has to decide the fate of one's body. For instance, it may be that in cases like The Fission Case, Ava's body ceases to exist in the freak accident, and yet it seems plausible that the ownership rights that would typically be described as being over her body still retain force.

Fourth, and finally, the forms of continuity mentioned above may correspond to the continuity of *patterns of appropriate concern* that we usually have for people over time. If the dementia patient is biologically and physically continuous with the signatory, and retains the signatory's capacity for consciousness, then it would make sense for the signatory's loved ones to continue to care about the dementia patient in many of the same ways that they would care about the signatory. Even if the signatory and the dementia patient are numerically distinct, given the abovementioned ways in which they are related, it would be absurd for the signatory's loved ones to completely withhold from the dementia patient the special concern that they once had for the signatory.

To sum up, the life of the dementia patient is in many ways, like an extension of the life of the signatory. And this suggests that an AD can still provide some moral reason regarding the person who signed it, to do what it says even when the person affected by that decision is not numerically identical to the person who signed the AD.

---

traits in real cases of severe dementia, see Caddell, L. S., & Clare, L. (2010) and Klein et al. (2003).

# 4. Continuity and different selves

I have argued that the numerical identity version of the someone else problem is not as poignant as writers such as Buchanan, DeGrazia, Dresser, and Kuhse seem to believe.

What about the different selves problem? Here, the form of the explanation we give regarding how the signatory's AD can have authority depends on how exactly we understand the different selves problem.

Let us assume for the sake of argument that in Anna's Directive, the signatory and the dementia patient are numerically identical. For simplicity, let's suppose that animalism is the correct theory of our identity over time, and that the signatory and the dementia patient are one and the same human organism—Anna.

Suppose that, following Das and Paul (2022), we accept a temporal parts ontology, and treat selves as temporal parts of human organisms. And suppose we let our terms 'the signatory' and 'the dementia patient' refer to different temporal parts of a single human organism (Anna). If we assume that selves (not human organisms) are the relevant objects of moral concern, then it seems we have just recreated the numerical identity version of the someone else problem. Since 'the signatory' and 'the patient' are different temporal parts of the organism, they (i.e., these two temporal parts) are not numerically identical.

But this means that we can respond to the different selves problem in much the same way we responded to the numerical identity problem. The same four points may be made in response to both problems: The fact of non-identity is not sufficient to rule out that the AD signed by the signatory provides a moral reason, one stemming from the signatory's expressed wishes, to honour the AD. The signatory and the dementia patient may still be psychologically connected, even if the psychological connections are too few to ensure psychological continuity. There is still physical, functional, and organizational continuity of the signatory's basic mental capacities, which establishes a prudential relation between the signatory and the dementia patient. Biological and bodily continuity are preserved, which ought to give the signatory some say over what is to be done regarding the dementia patient. (Here selves, not human organisms, are treated as the bearers of bodily rights.) Finally, the special concern (for example expressed by loved ones) for the signatory extends, to some degree, to the dementia patient.

However, suppose we treat human organisms as the relevant units of moral concern. In that case, the different selves problem is that the individual (Anna) has changed her core preferences, and the preferences she has at t2 conflict with those she has at t1. The question then is why Anna's preference at t1 should still matter at t2 when she no longer has that preference.

In Section 3, I acknowledged that it can seem intuitive that a person's past preferences do not provide a moral reason regarding that person to act against her current preferences in cases of radical psychological change. I considered the real case of an amnesiac (Richard) who no longer desires to live with his family members. The fact that he used to desire that does not seem to provide a moral reason regarding Richard to force him into his old life.

However, there are examples that seem to support the claim that an individual's past preferences can matter morally regarding what we do now, even if she no longer has those preferences.[10]

Consider Parfit's example of the Russian nobleman who currently stands to inherit vast estates in several years and has a core preference for giving away this inheritance (1984, p. 327). Since he knows that his preferences may change with time, he signs a legally binding document to give away his estates, which can be revoked only with his wife's consent. He then asks his wife to promise not to consent if he later changes his mind and wants her to revoke the document. Regarding his current values as essential to him, he tells his wife that if his values change, he wants her "to regard your husband then, not as me, the man who asks you for this promise," but as a different self.

Suppose the nobleman's wife promises not to revoke the document, and the nobleman later changes his core preferences and asks her to revoke it. One would then expect his wife to feel, at the very least, *conflicted*. This suggests that the nobleman's past preferences matter, even though he no longer holds them when his wife can decide whether to satisfy or frustrate them.

Parfit's case of the Russian nobleman suggests that a person's past preference provide moral reasons regarding that person to act against their current preferences. But the case of Richard the amnesiac suggest that they don't. My suggestion is that past preferences provide moral reasons in both cases, but that in the second case it is practically impossible to satisfy the past preference, and this explains the apparent absence of any moral reason.

This becomes clearer when we closely consider the specific content of Richard's past preference. It seems plausible that Richard's past preference to continue living with his family *would not be satisfied* by forcing him to do so. This is because, pre-

[10] One could claim that in fact the dementia patient at t2 still has the core preferences held by the signatory at t1 but is unable to express them. The problem is that there does not appear to be any reason to believe this in general. Another possible response is that the change in core preferences over time was due to an aberrant causal process outside the control of the individual, and that therefore the preferences of the dementia patient are not the individual's "true preferences". However, it is unclear why the type of causal process should be thought to matter morally. Individuals change their preferences for many different reasons, and usually these changes involve some type of external cause that is beyond their control.

sumably, Richard's past preference that he remain together with his family is *not unconditional*. He would not want to be forced into a life with his family members if that life completely lacked the emotional and cognitive relations that had formed over their previous time together. Such a life would likely leave both Richard and his family utterly miserable.[11]

However, it may still seem intuitive in many cases that a person's past preferences do not provide a moral reason to act against their current preferences, even when we *could* satisfy their past preferences. I will now offer a diagnosis of what is going on in such cases.

Here, let us focus on Anna's Directive. My suggestion is that, at the time when the caretakers must make their decision regarding whether to continue life support for the dementia patient, the preferences of the dementia patient constitute the only moral consideration that is *salient* to the caretakers. At t2, Anna's earlier self that preferred not to lack awareness of her loved ones, is, in effect, out of mind, and hence, her t1-preferences seem not to matter.

To test this hypothesis, we can imagine a case in which the dementia is miraculously reversed, bringing the signatory back into focus. Consider *Anna's Return*. This case is just like Anna's Directive but with the following additional detail: After Anna contracts a life-threatening case of pneumonia, her caregivers see no good reason to withhold life-support and decide to help Anna continue to live. However, after several weeks, Anna regains a remarkable level of cognitive function. She again recognizes her children, recalls her earlier life, and resumes her role as a reflective, articulate individual. No one could reasonably have expected Anna's dementia to reverse. Her recovery is truly a miracle.

After recovering her memories, it seems Anna should be happy to be alive. And yet it would be appropriate for Anna to also feel indignation toward the caretakers for ignoring her directive. This suggests that they had a moral reason regarding Anna to honour that directive. The existence of this reason is not plausibly explained merely by Anna's recovery. For we have stipulated that there was no reason to believe she *would* recover.

Instead, it seems that when Anna recovers, her previous wishes, which were disregarded, once again become salient—they are brought back into focus. If those past preferences did not ground any moral reason, then Anna's indignance after recovering would be completely groundless. This suggests that if the caretaker's give Anna's past preferences no weight, this is because Anna's past self simply is not salient

[11] A more interesting question is whether in this case there is a moral reason regarding Richard to force a reversal of his amnesia. Suppose we could now slip him a magic pill that would bring back his memories and emotional connections to his family. Would there be a moral reason to do that, even without his present consent? I think there is such a reason, even if it would not be decisive.

at the time of deliberations. But this lack of salience does not, by itself, have any moral relevance.

# 5. Conclusion

Advance directives in cases of severe dementia are vulnerable to a deep philosophical worry: that the authority they claim appears to be undermined when the person who signs the directive and the later dementia patient are, in a morally significant sense, "someone else." I have distinguished two versions of this worry. The first, the numerical identity version, assumes that the signatory and the dementia patient are numerically different people. The second, the different selves version, grants numerical identity but claims that the dementia patient has become a different self—one with different core preferences, character traits, and values.

Against both versions, I have argued that we can still make sense of the authority of an advance directive. In response to the numerical identity version, I drew on Parfit's fission examples to argue that numerical identity is not necessary for survival, and not necessary for the authority of prudential or moral reasons across time. Even if identity is lost, the relevant forms of continuity—psychological, biological, and narrative—can ground reason-giving relations between the signatory and the later patient.

In response to the different selves version, I appealed to the idea that past preferences—even those that no longer define the current self—can retain normative significance. Borrowing from Parfit's Russian nobleman and from a hypothetical case involving reversible dementia, I suggested that a person's past preferences can ground a moral reason, regarding that person to do as she preferred, even when the person currently has conflicting preferences.

The conclusion, then, is modest but important. The authority of advance directives is not absolute, but neither is it defeated by psychological change. Whether or not the dementia patient is "someone else" in a metaphysical or psychological sense, we can still have morally relevant reasons to honor the directive. The challenge is not to prove that the signatory and the dementia patient are the same person, or even the same self. It is to recognize and weigh the normative significance of past preferences against that of present preferences—and to do so in a way that respects both the person who is and the person who was.

# References

Buchanan, A. E. (1988). Advance directives and the personal identity problem. *Philosophy & Public Affairs, 17*(4), 277–302.

Caddell, L. S., & Clare, L. (2010). The impact of dementia on self and identity: A systematic review. *Clinical Psychology Review, 30*(1), 113–126.

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* (pp. 175–196). Routledge.

DeGrazia, D. (2000). Advance Directives, Dementia, and 'the Someone Else Problem'. *Bioethics*, 14(2), 90–103. https://doi.org/10.1111/1467-8519.00184.

DeGrazia, D. (2005). Human Identity and Bioethics. Cambridge University Press.

DeMarco, J. P., & LiPuma, S. H. (2016). *Dementia, advance directives, and discontinuity of personality*. *Cambridge Quarterly of Healthcare Ethics, 25*(4), 674–685. https://doi.org/10.1017/S0963180116000396.

Dresser, R. (1995). Dworkin on dementia: Elegant theory, questionable policy. *The Hastings Center Report*, 25(6), 32–38. https://doi.org/10.2307/3562659.

Dworkin, R. (1993). *Life's Dominion: An Argument About Abortion, Euthanasia, and Individual Freedom*. Alfred A. Knopf.

Klein, S. B., Cosmides, L., & Costabile, K. A. (2003). Preserved knowledge of self in a case of Alzheimer's dementia. *Social Cognition, 21*(2), 157–165.

Markowitsch, H. J., & Staniloiu, A. (2025). *Episodic-autobiographical memory and functional amnesia*. *Translational Neuroscience, 16*, 20250387. https://doi.org/10.1515/tnsci-2025-0387.

McMahan, J. (2002). *The ethics of killing: Problems at the margins of life*. Oxford University Press.

Olson, E. T. (1997). *The human animal: Personal identity without psychology*. Oxford University Press.

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

Persad, G. (2019). *Authority without identity: Defending advance directives via the someone else problem*. Journal of Medical Ethics, 45(2), 125–130. https://doi.org/10.1136/medethics-2018-104971

Van Inwagen, P. (1990). *Material beings*. Cornell University Press.

Walsh, E. (2020). Cognitive transformation, dementia, and the moral weight of advance directives. *The American Journal of Bioethics, 20*(8), 54-64. https://doi.org/10.1080/15265161.2020.1781955.

Wiggins, D. (1967). *Identity and Spatio-Temporal Continuity*. Oxford: Basil Blackwell.