22

PERSONAL IDENTITY AND BRAIN IDENTITY

Georg Northoff and Nils-Frederic Wagner

Introduction

Say you are looking at an old picture of your high school graduation and recognize yourself as the teenager with the funny haircut. Perhaps, from today's perspective, the haircut is not all that fashionable anymore, but nonetheless you are certain that the person in the picture is you. But what makes it true that you today and the teenager in the picture are identical—or one and the same person over time? This is a question of *diachronic* personal identity. In order to answer these kinds of questions, we must know the *criterion* of personal identity. In other words, we want to know what the necessary and sufficient conditions are that account for a person persisting from one time to another.

Numerical and Qualitative Identity

When philosophers debate personal identity, they are mostly concerned with *numerical* identity, the relationship that can hold only between a thing and itself. Numerical identity appears to require absolute or total qualitative identity. The puzzle is how, despite qualitative changes, a person still remains one and the same and persists through time and change. For example, say Jane radically changed in her personality traits, as well as in her appearance, due to a religious conversion. These changes, however, do not make Jane cease to exist altogether; they rather alter her *qualitative* identity. In questions about numerical identity, we look at two names or descriptions and ask whether these refer to one and the same person at different times or to different persons. The basis for being distinctively concerned about one's own future in a way that is inevitably different from how we are concerned about someone else's future is widely believed to be grounded in numerical identity. However much you will change, you shall still exist, if there will be someone living who will be numerically identical to you.

It goes without saying that it is impossible for a single person at different times to be identical to themselves in a strict numerical sense, especially if taking into account that the human body's cells are constantly replaced and the brain cell connections and chemistry are frequently changing minute to minute. But this isn't the kind of identity that raises questions concerning one's own survival over time. It is rather what David Wiggins (1967) calls the "conditions of persistence and survival through change". An understanding of personal identity that persists

through change is, both from a pretheoretical point of view and after conceptual analysis, more compelling than the appeal to strict numerical identity (wherein a person is somehow immutable and unchanging). So when we think of a person remaining one and the same over time, we don't usually think that this prevents that person from undergoing qualitative changes. It would be absurd to claim that you aren't the person you were yesterday because your hair grew a tiny bit overnight. For this reason, accounts of personal identity must allow for persons to change in their qualitative features and nonetheless account for their persistence through time.

It is widely held (although some philosophers disagree) that questions about numerical identity must always be *determinate*, that is, whether you will exist some time in the future must always have a clear-cut 'yes/no' answer (Evans, 1985).

Criteria of Personal Identity

In what follows, we give a swift overview of the most widely held criteria of personal identity and point to some of their implications and problems.

Identity, Reductionism and Nonreductionism

According to *reductionist* theories, personal identity is reducible to more particular facts about persons and their bodies. The approach is to describe a particular relation (call it Relation *R*) that accounts for how person X is identical to a later existing person Y by virtue of X and Y being *R*-related. In other words: X is one and the same person as Y, if and only if X stands in Relation *R* to Y. In principle, Relation *R* is believed to be empirically observable. However, there is major disagreement about what Relation *R* consists in. Philosophers disagree about which particular ingredients determine the relation that constitutes personal identity. In the contemporary debate, most philosophers hold one or another form of a reductionist account. According to physical/biological reductionism, a person is identical over time so long as they remain the same living organism. For example, you are identical to the fetus you once were because a biological trajectory of your current body can be traced back to the fetus that once was in your mother's womb. A more widespread version of reductionism is psychological reductionism, according to which a person is identical over time so long as their different temporal parts are connected through psychological continuity. For example, you are identical to the person in that high school graduation picture because your current mental states (memories, desires, plans, intentions etc.) are connected through overlapping chains of psychological connectedness. We'll discuss seminal versions of both biological and psychological reductionism in more detail in the following section.

In contrast to reductionist theories of personal identity, *nonreductionists* believe that personal identity is not reducible to more particular facts about persons and their brains and bodies but rather consists in a nonanalyzable, or *simple*, 'further fact'; for example, an indivisible entity that eludes further analysis by being all by itself the necessary and sufficient condition for personal identity over time. Derek Parfit (although himself endorsing a reductionist criterion) describes the notion of a further fact as "separately existing entities, distinct from our brains and bodies, and our experience" (Parfit, 1984, 445). Nonreductionists thus claim that personal identity consists in a special ontological fact, a Cartesian Ego (going back to Descartes' substance dualism) or a soul; or stated in a less antiquated way, the view is that personal identity consists of some mental entity that is not reducible to neural mechanisms in the human brain. In the contemporary discussion in philosophy of mind, few philosophers advocate for nonreductionist accounts of personal identity because those accounts are, at least to the majority of philosophers, metaphysically

contentious. It is argued that nonreductionists in the debate on personal identity take an obscure metaphysical belief and inflate it into a conceptual core conviction.

Psychological Continuity and Animalism

Granting the aforementioned concerns about nonreductionism, we will not further elaborate on these accounts. Instead, we focus on the two most paradigmatic reductionist accounts of personal identity: seminal versions of the *psychological continuity* theory and *animalism*.

According to the psychological continuity criterion of personal identity, X and Y are one and the same person at different points in time if and only if X stands in a *psychological continuity* relation to Y. You are the same person in the future (or past) as you are now if your current beliefs, memories, preferences and so forth are linked by a chain of overlapping psychological connections. Among philosophers who advocate for psychological approaches to personal identity, there is dispute over several issues: What mental features need to be inherited? What is the cause of psychological continuity, and what are its characteristics? Must it be realized in some kind of brain continuity, or will 'any cause' do? The 'any cause' discussion is based on the counterfactual idea that personal identity that is realized by psychological continuity would still hold, even if this continuity would no longer be instantiated in a brain. For example, would psychological continuity still determine identity if the personality was instantiated in a computer program? That is to say, what happens if the psychological relations that define personal identity get replicated and instantiated in a nonbiological entity or in more than one biological entity? Since psychological continuity could in principle divide, another issue is whether a 'nonbranching clause' is needed to ensure that psychological continuity holds to only one future person. It has been argued that the logical possibility of psychological continuity splitting into more than one successor evokes the need of blocking such scenarios by implementing a one–one clause into the criterion (we'll come back to this point in the section on fission cases).

According to *animalism*, you are the same being in the future (or past) as you are now (or have been earlier), as long as you are the same living biological organism. A human animal, or for that matter any organism, persists as long as its capacity to direct those vital functions that keep it biologically alive are not disrupted. If X is an animal at time *t* and Y exists at time $t^{\star}$, X and Y are identical if and only if the vital functions that Y has at $t^{\star}$ are causally continuous (without splitting into two or more successors) with those that X has at t. Presumably this will be the case only if Y is an animal at $t^{\star}$. So anything that is an animal at one time will always be an animal, and identity between an animal at one time and at another time is maintained when those vital functions are causally continuous. For animalists, psychological features have no bearing on personal identity because human animals go through periods without having any mental functions.

The difference between these two criteria becomes apparent when considering cases at the margins of life. A fetus has no psychological features, and thus according to the psychological continuity view, no person is diachronically identical to a fetus. Whereas for the animalist, a fetus and the person it later becomes are identical by virtue of being the same (single) living organism. The same holds for other cases in which human organisms lose their mental capacities but remain biologically alive.

Thought Experiments

Hypotheticals such as John Locke's famous 'Prince and the Cobbler' are still widely discussed in the metaphysical debate on personal identity (Weinberg, 2011). Locke asks what would happen if the soul of a prince, carrying with it the consciousness of the prince's life, were to enter the

body of a cobbler. Locke suggests that as soon as the cobbler's body is invaded by the prince's soul, the cobbler would be the same *person* as the prince, accountable only for the prince's actions. But who would say the prince was, in Locke's term, the same 'man' or human animal? With this thought experiment, Locke suggests that persons, unlike human animals, are only contingently connected to their bodies. Locke further maintains that what constitutes a person, and moreover the *same* person, is consciousness, the awareness of one's thoughts and actions: "Nothing but consciousness can unite remote existences into the same person" (Locke, 1694/1975, 464). Referring to a man he had met who believed his soul had been the soul of Socrates, Locke asks, "If the man truly were Socrates in a previous life, why doesn't he remember any of Socrates' thoughts or actions?" Locke even goes so far as to say that if your little finger were cut off and consciousness should happen to go along with it, leaving the rest of the body, then that little finger would be the person—the same person that was, just before, identified with the whole body (Locke, 1694/1975, 459–460). On these grounds, Locke and his modern-day successors establish that wherever your mental life goes, you as a person follow.

Brain Transplants

In the spirit of Locke's 'Prince and the Cobbler', brain transplant thought experiments figure prominently in the personal identity literature. Recently the idea has entered the public debate as a science fiction prospect for future medical use. The story goes something like this: imagine someone's brain (or their cerebrum as the seat of their distinct psychology) is transplanted into someone else's empty skull (or their brain with a removed cerebrum). Then we are asked to ponder: who is the person that wakes up after the operation has been performed; is the resulting person identical to the 'brain donor' or to the 'body donor', or is it an altogether different person? Ever since Sydney Shoemaker (1963) introduced these sorts of imagined cases into the modern debate, they are frequently presented as support for psychological continuity theories and seen as troublesome for competing views. Particularly, they are fairly often regarded as more or less decisive evidence against animalist takes on personal identity. However, there are at least two lines of reasoning against this interpretation. Advocates of bodily continuity theories such as Bernard Williams (1970) have claimed that a variant of brain transplants, sometimes described as 'body swapping', actually works in favor of bodily continuity views. By slightly altering the story, Williams has shown that our intuitions as to who the resulting person will be can easily tilt. Modern-day animalists such as Eric Olson (1997) are allies in this interpretation of brain transplants. Paul Snowdon (2014) has recently presented further arguments questioning the alleged support for psychological continuity theories gathered from hypothetical brain transplants. Another, more general line of criticism comes from Kathleen Wilkes (1988), who claims that due to an inevitable lack of detail in the description of hypothetical scenarios, conclusions drawn from there often lead to a false reliance on predictions about how our concept of personal identity would apply in the imagined case. On this view, brain transplant thought experiments do not support psychological continuity theories but simply track intuitions that have no bearing on our concept of personal identity.

Fission Cases

Despite their initial appeal, psychological continuity theories share a severe problem. Unlike identity, psychological continuity is not necessarily a one–one relation; that is, psychological continuity does not necessarily hold only to one future/past person but can hold to many. For example, fission scenarios, either based on purely hypothetical cases or based on brain bisection

(corpus callosotomy), as put forward, among others, by Thomas Nagel, show that psychological continuity does not follow the logic of an identity relation (Nagel, 1971). Cutting the connection between the two hemispheres is used as a treatment to restrict epileptic seizures to one hemisphere. Callostomized patients develop unconnected streams of consciousness in both visual fields, resulting in sometimes contradictory descriptions and actions. It has been reported that patients simultaneously tried to embrace their partners with one hand and push them away with the other (Sperry, 1966). It is possible in principle, then, and in accordance with empirical evidence, that psychological continuity can divide when the physical brain is divided, and thus, multiple persons can be psychologically continuous. As David Lewis and others pointed out, identity is necessarily a one–one relation that can by definition only hold to itself, whereas psychological continuity is only contingently a one–one relation and may become one–many (Lewis, 1976). Imagine that a split-brain patient's two hemispheres were transplanted into two different heads. Both resulting people would stand in a psychological continuity relation to the original person. Therefore, as Bernard Williams argued, psychological continuity is unable to meet the metaphysical requirements of a criterion for personal identity unless a nonbranching clause is added that ensures that psychological continuity can only be a one–one relation (Williams, 1970). Nevertheless, the addition of such a nonbranching clause is not fully convincing either, because, as Derek Parfit claimed, a nonbranching clause has no impact on the intrinsic features of psychological continuity and is therefore unable to explain the difference between the importance that we attach to identity and the relative unimportance of psychological continuity if it takes a branching form. So, if you are psychologically continuous with only one person, identity holds, but as soon as someone else is also psychologically continuous with the original person, identity vanishes. In any case, identity could not be sustained over time, since two persons would diverge in their experiences and their psychological contents before very long. Parfit famously concluded that "identity is not what matters" (Parfit, 1984).

The Brain Criterion

Given the aforementioned difficulties facing psychological continuity views—some of which result from detaching psychological continuity from brain continuity—and the counterintuitive implications of animalism, another criterion of personal identity has gained some recent support. According to philosophers Thomas Nagel and Jeff McMahan, personal identity is secured through *brain identity* (Nagel, 1986; McMahan, 2002). This is an attempt to base personal identity on solid empirical grounds, and it opens up new possibilities of cohering philosophical accounts of personal identity with neuroscientific evidence. This notion of personal identity has recently been discussed in the context of neuroscience (Northoff, 2001; 2004). More specifically, manipulation of the brain may change not only its neuronal functions but also its psychological and mental functions. For instance, implantation of tissue or electrodes into the brain may change the neuronal underpinnings of psychological features that are considered crucial in psychological continuity views. Does this mean that brain manipulation entails manipulation of personal identity? If so, one would assume brain identity to be a necessary and/or sufficient criterion of personal identity.

The recent insight into the empirical functioning of the brain raises a whole new field for the philosophical discussion of personal identity: the relationship between personal identity and the brain. Since the brain cannot be understood in a purely conceptual manner but must rather be explored empirically, this new field must necessarily link conceptual and empirical domains and thus be truly neurophilosophical in the genuine sense of the term. We'll now discuss Nagel's

and McMahan's brain view of personal identity in some detail; subsequently we offer some neurophilosophical reflections on these two approaches.

You Are Your Brain: Thomas Nagel

Nagel (1986) considers the brain as both a necessary and sufficient criterion for personal identity, stating that "I am my brain" (Nagel, 1986, 64–5, 69). Nagel thinks that questions of personal identity are determinant—necessarily yielding a 'yes/no' answer. Personal identity, for Nagel, is nonconventional in the sense that it does not imply its own necessary and sufficient conditions, thus containing an 'empty position' that must be filled by an 'additional fact' (Nagel, 1986, 71). Nagel compares the term 'identity' with the term 'gold', saying that, before the chemical formula for 'gold' was discovered, the term 'gold' contained an 'empty position' that was later filled by an 'additional fact': the chemical formula for gold.

The same goes for personal identity and the brain. We currently have no idea about the 'additional fact' that might potentially fill the 'empty position' in personal identity. According to Nagel, the 'additional fact' in the case of personal identity must bridge the gap between the subjective experience of the person, that is, its first-person perspective and its 'I' on the one hand, and the objective, necessary structures, that is, its third-person perspective and its physical body on the other. Since the brain might eventually bridge this gap between subjective experience and objective structures, it may be considered a suitable candidate to fill the 'empty position'. The subjective component of personal identity comes into play as introspective evidence for persistence over time via some form of psychological continuity. The objective dimension of personal identity comes into play through the brain, which, being a physical organ, has purely objective features that enable us to track a person's identity without relying on fallible introspection. But at the same time, the brain accounts for subjectivity.

Subjectivity and Objectivity

Nagel thinks that personal identity cannot be fully "understood through an examination of my first-person concept of self, apart from the more general concept of 'someone' of which it is the essence" (Nagel, 1986, 35). Nagel further believes that personal identity cannot be defined *a priori*. A 'subject of consciousness' must be able to self-identify without external observation; but these identifications must correspond to those that can be made on the basis of external observation. Thus, there are both first- and third-personal features of the self and its identity over time. There must be a notion of objectivity, Nagel says, which applies to the self, for it is clear that the idea of a mistake with regard to my own personal identity makes sense. The objectivity underlying this distinction must be understood as objectivity with regard to something subjective. The question is how the idea of the same subject can meet the conditions of objectivity appropriate for a psychological concept. It is subjective (not merely biological) but, at the same time, admits the distinction between correct and incorrect self-identification.

The Brain as a Bridge Between Subjectivity and Objectivity

What could perform the function of a special type of material substance that also has irreducible subjective features? Subjects of experience are not like anything else. While they do have observable properties, the most important thing about them is that they are subjective, and it is their subjective mental properties that must be accounted for. Nagel suggests that the concept of the self is open to 'objective completion' provided something can be found that straddles the

subjective/objective gap. Something whose objective persistence is among the necessary conditions of personal identity is needed, but only if this objectively describable referent is in a strong sense the basis for those subjective features that typify the persistent self. It must refer to something essentially subjective, often identifiable nonobservationally in the first-person perspective and observationally in the third-person perspective, and something that is the persisting locus of mental states and the vehicle for carrying forward familiar psychological continuities when they occur. Nagel thinks that the brain is the most plausible contender to fulfill both the objective and subjective demands of personal identity. I could lose everything but my functioning brain and would still be me.

Accordingly, the brain might be considered an 'additional fact' in Nagel's sense. On the one hand, the brain must be considered the necessary foundation for the possibility of subjective experience, since without a brain we remain unable to experience anything. On the other hand, the brain is the carrier of psychophysiological processes that remain essential for regulation and maintenance of the body. In contrast to other organs like, for example, the liver or kidney, the loss of the brain is accompanied by the loss of personal identity. The brain must subsequently be regarded as a necessary and sufficient condition for personal identity. However, Nagel concedes, due to our current lack of empirical knowledge, this assumption must be considered a preliminary hypothesis.

'Physico-Mental Intimacy'

As we have said, Nagel claims that the brain bridges the gap between the subjective experience of mental states and psychophysiological states. The brain can therefore not be considered a purely physical organ since, for Nagel (1986), mental states and subjective experience cannot be reduced to physical properties (57, 74). In addition to physical properties, the brain must therefore be characterized by mental properties. Nagel speaks of a so-called physico-mental intimacy (in an ontological sense) as an "apparent intimacy between the mental and its physical conditions" (Nagel, 1986, 20).

Due to these mental properties, the brain shows a special kind of 'insideness' that accounts for its foundational character for subjective experience: "It [the brain] can be dissected, but it also has the kind of inside that can't be exposed to dissection. There's something it is like from the inside to taste chocolate because there's something it's like from the inside to have your brain in that condition that is produced when you eat a chocolate bar" (Nagel, 1987, 34–35). Accordingly, the brain can be described mentally and physically, and both may be traced back to what Nagel calls a "fundamental essence" (Nagel, 1979, 199). This fundamental essence can be defined by complex forms of organization and combinations of matter—an "unusual chemical and physiological structure"—, which shows both proto-physical and proto-mental properties (Nagel, 1979, 201). However, neither the exact definition of both kinds of properties nor their relation is clearly determined. Thus, both important conceptual and empirical details are in need of further clarification.

Embodied Minds: Jeff McMahan

Another recent version of the brain criterion is the view that Jeff McMahan (2002) dubbed the 'Embodied Mind' account. McMahan argues that the continuity of parts of the brain that generate and sustain consciousness is both necessary and sufficient for personal identity over time. In order to survive as the same person over time, the capacity of a person's consciousness must be realized by the continuity of the same brain.

Minds

McMahan holds that a mind is individuated by reference to its physical embodiment, just as an individual mental state is. A particular memory, for example, continues to exist only if the tissues of the brain in which it is realized continue to exist in a potentially functional state. Likewise, a particular mind continues to exist only if enough of the brain in which it is realized continues to exist in a functional or potentially functional state. This neatly explains how minds persist. If a single mind has hitherto been realized in certain regions of a single brain, then, the undivided survival and continued, self-sufficient, functional integrity of those specific regions is both a necessary and a sufficient condition of the continued existence of the same mind (for further details see McMahan, 2002, 66 ff).

McMahan's take on personal identity is practical; he thinks that the basis for an individual's egoistic concern about the future is the physical and functional continuity of enough of those areas of the individual's brain in which consciousness is realized. The person persists as long as the capacity to support consciousness is preserved. Now, usually the functional continuity of these areas of the brain entails broad psychological continuity. But in the very earliest phases of an individual's life and in some instances near the end, the same mind or consciousness persists in the absence of any degree of psychological connectedness from day to day. McMahan's criterion stresses the survival of one's basic psychological capacities, in particular the capacity for consciousness; it does, however, not require the continuity of any of the particular contents of one's mental life. For example, one continues to exist throughout the progress of Alzheimer's disease until the disease destroys one's capacity for consciousness (McMahan, 2002, 71).

Brain Continuity

McMahan distinguishes among three types of continuity of the brain. His idea of *physical continuity* of the brain is applicable in either of two ways. It could involve the "continued existence of the same constituent matter" of the brain that generates and sustains consciousness or "the gradual, incremental replacement of the constituent matter of the brain over time" (McMahan, 2002, 68). In the latter sense, the brain, like most other organs of the human body, can survive gradual cellular turnover, which, on his view, is congruent with physical continuity. Thus, the core principle in preserving physical continuity over time is that there is sufficient integration between the old and new matter, which rules out in advance the compatibility of rapid replacement with physical continuity.

Closely related to physical continuity is what McMahan calls *functional continuity*. This involves roughly the continuity of basic psychological capacities of the brain—in particular, the brain's capacity for consciousness, which is sufficient for minimal functional continuity. A third and obviously less important type of continuity for McMahan is *organizational continuity*. It involves the continuity of the various tissues of the brain that underlie the connections among the distinctive features of one's psychology. These include the connection between an earlier experience and a later memory of it.

One would think that functional and organizational continuity presuppose physical continuity, but McMahan assumes that, as long as certain functions or patterns of organization are preserved, there will be functional or organizational continuity even if the relevant functions or patterns of organization are not preserved in the same matter.

So McMahan's criterion of personal identity and egoistic concern is physical and minimal functional continuity of the brain; more specifically, enough of the relevant areas of the brain in which consciousness is generated and sustained—to be capable of preserving the capacity for

consciousness (McMahan, 2002, 67–69). I am the same entity today as I was yesterday because the same brain supports the same capacity for consciousness it supported yesterday. The continuity of the distinctive features of my mental life (memories, beliefs etc.) is incidental to my survival; what is important is physical continuity of the brain.

Reducing personal identity to physical and minimal functional continuity of the brain distinguishes McMahan's view from other accounts, such as psychological continuity views that include organizational continuity in the criterion of personal identity. McMahan thinks that the continuing capacity for consciousness is a sufficient basis for egoistic concern about one's own future and should, therefore, be a sufficient basis for personal identity, other things being equal. Since McMahan holds that a person begins to exist with the onset of (or the capacity for) consciousness in their organism, fetuses and patients with brain damage potentially fall under the category of beings with future interests as well. McMahan tries to make this work by emphasizing that unlike identity, the basis for egoistic concern can be present in degrees. Defending a morality of respect, McMahan makes the distinction between the badness of death and the wrongness of killing when it comes to beings with little psychological life like fetuses. The badness of death for persons themselves can vary dramatically from one person to another—young, old, gifted, ungifted and so on. The wrongness of killing someone doesn't depend on their egoistic concern about their future. Killing a person is equally wrong in each case because it violates respect for the person.

Neurophilosophical Reflections

After having considered the two most prominent views advocating brain identity as the criterion for personal identity over time, we are now in a position to critically reflect on these views from a neurophilosophical perspective. Subsequently, we sketch in broad strokes a few additions to the existing literature on brain identity that might be helpful to tie up some of the loose ends.

Is Personal Identity an 'Additional Fact'?

Nagel assumes an 'additional fact' about the brain in his concept of personal identity. This 'additional fact' should be empirically accessible. However, the psychological data show no hints or indications for such an 'additional fact.' Nobody experiences anything but their personality, which they regard as the sum of their psychological functions. In contrast, we do not define (or experience) our personalities (and personal identities) by any particular property or fact which could be regarded as equivalent to Nagel's 'additional fact'.

Nagel could argue that this 'additional fact' is epistemically not directly accessible and somehow unknowable. This leaves him with the following three options: (1) The 'additional fact' might not be accessible at all, neither first-personally nor third-personally. In this case, the 'additional fact' remains in principle hidden from us. One might consequently consider the 'additional fact' as rather mysterious; accordingly, this option does not seem to be very attractive. (2) The 'additional fact' might be accessible but merely third-personally. Then it should be detectable in neuroscientific investigations that rely on the third-person perspective. However, such an 'additional fact' has not yet been detected in neuroscientific investigations of the brain. This option thus remains empirically rather implausible. (3) The 'additional fact' might be accessible merely first-personally, though in a disguised or indirect form. In this case, the 'additional fact' might not be accessible as a 'fact' but rather as a particular type of state as distinguished from neuronal states. Instead of looking for an 'additional fact', one should then aim at revealing an 'additional state' and its relation to the brain. What could this 'additional state' be? Is it a mental

state? If mental states do account for the 'additional state' as a disguised form of the 'additional fact', and thus the brain, mental states must then reflect the access to one's own brain from the first-person perspective. We may perceive our own brain states thus not as brain states but as mental states. Our brain as the 'additional fact' can thus be accessed only indirectly via mental states as 'additional states'. This probably comes closest to what Nagel has in mind (or rather in his brain?).

Conversely, we remain unable to access our own brain as a brain from the first-person perspective. This epistemic inability of our own brain to access itself directly as a brain can be called an 'autoepistemic limitation' (Northoff, 2004). This 'autoepistemic limitation' may be subserved by specific principles of functional brain organization that prevent the brain from directly perceiving itself as a brain. These principles of functional brain organization may thus fulfil the same role as the chemical formula for gold in Nagel's example. Accordingly, 'autoepistemic limitation' might account for what Nagel means by 'additional fact', although his formulation 'I am my brain' should then be rephrased as 'I am my brain, but due to autoepistemic limitation, I remain unable to directly access myself as a brain'. Moreover, an investigation of the empirical mechanisms underlying 'autoepistemic limitation' requires direct relationships between empirical functions and epistemic abilities/inabilities, that is, a so-called epistemic–empirical relationship (Northoff, 2004). This in turn makes the development of 'neuroepistemology' (Northoff, 2004) as an 'epistemology on a neurological basis' (Kuhlenbeck, 1965, 137) necessary.

'World–Brain Relation' as 'Additional Fact'?

The mere physical continuity of the brain is by itself not sufficient to account for personal identity over time. Nagel assumes an 'additional fact,' and McMahan seems to take the brain only as the seat of consciousness. How can we determine the 'additional fact' in a nonmental way without reverting to any kind of metaphysics involving mental features like consciousness? The proponent of the brain criterion may argue that the presupposition of the brain as a mere placeholder for mental features deflates the importance of the brain. We may do better by considering the brain itself. But then we are again confronted with the problem that the brain itself does not seem sufficient for personal identity.

All criteria that rely on the brain presuppose the brain as an isolated organ (or just as the seat of consciousness). However, there are empirical data that suggest that the brain and its own spontaneous activity, for example, its intrinsic activity, are strongly dependent upon the respective environmental context. For instance, a study by Duncan et al. (2015) showed that traumatic life events in early childhood impact the brain's spontaneous activity in adulthood: the higher the degree of early traumatic childhood life events (i.e., either a greater number of traumatic events or events that are more traumatic), the higher the degree of spatiotemporal disorder (e.g., entropy, in the brain's spontaneous activity in adulthood). This makes it clear that the brain is not an isolated organ encapsulated in the skull that communicates with the environment only through the body. For instance, the brain can shift the phase onset of its fluctuations (in different frequency ranges) in accordance with the onset of tones in the external environment (see chapter 20 in Northoff, 2014b, for details). That is apparent when we listen to music and swing our perceptions and movements in tune with the rhythm. Such direct coupling between brain and environment has also been described as 'active sensing' or the brain's 'sixth sense' (van Atteveldt et al., 2014). This and other examples (Northoff, 2014b; 2016) make it clear that the brain is a highly context-dependent organ that stands in direct relation to the respective environment, with the latter impacting, modulating and sculpting the former. Taken in this sense, the brain can no longer be considered an isolated organ but a relational organ that can

be characterized by what we describe as a world–brain relation or 'environment–brain unity' (Northoff, 2014b).

Why is the characterization of the brain as 'environment–brain unity' relevant for determining personal identity? The relation in which the brain stands to its environment is apparently an intrinsic and thus defining feature of the brain as brain. If so, one may assume that this world–brain relation may account for the 'additional fact' Nagel was searching for when he conceived of the brain as seat of personal identity. One may then want to rephrase Nagel as saying, 'I am my world–brain relation', rather than, 'I am my brain'.

The assumption of physical continuity as suggested by McMahan may then need to be replaced by relational continuity: as long as I stand in a continuous relation to the world, I remain identical even if my brain and its physical continuity change. Such relational continuity presupposes a different ontology, however. Instead of a property-based ontology, one may then want to suppose a relational ontology that comes close to what Alfred North Whitehead described as 'process ontology' (Whitehead, 1929/1978). Process ontology takes the continuous change and dynamics, and thus the flux of being, as the most fundamental unit of existence and reality. Ontology is here essentially dynamic with continuous change across time and space. That stands in opposition to property-based ontology in which specific properties (like physical or mental properties) that are static and nonchangeable are assumed as the basic units of existence and reality.

Personal Identity as 'Physico-Mental Intimacy'?

Patients with brain implants report a close interaction between their mental states and the physical substitutes, that is, the cells or the stimulating electrode (Northoff, 2001). For example, they describe the feeling that they could influence the brain by their psychological and mental states. Nagel's assumption of 'physico-mental intimacy' might thus be considered as empirically plausible. The subjective experience of being able to influence mental states both via one's own brain tissue and through implants would not be possible without some kind of 'physico-mental intimacy'—subjectively accessible via the first-person perspective.

Does this epistemic characterization of 'physico-mental intimacy', however, justify the inference of physical and mental properties in an ontological sense? Such an epistemic-ontological inference is, for example, reflected in Nagel when he infers from the epistemic description of a special 'insideness' to underlying ontological properties. This type of inference might be called an 'epistemic-ontological fallacy' (Northoff, 2004b) in which one falsely infers from epistemic or knowable characteristics to ontological properties or the essence of being.

However, even if one allows for epistemic-ontological inferences, the ontological assumption of mental properties of the brain, as distinguished from its physical properties, cannot easily be justified: though mental states can be experienced in the first-person perspective, their experience cannot be directly linked to one's own brain because of 'autoepistemic limitation'. If, however, mental states cannot be directly linked to one's own brain, any type of epistemic-ontological inference from mental states to mental brain properties remains impossible. Nagel could, however, argue that if mental brain properties cannot be inferred from the experience of mental states in the first-person perspective, they may at least be inferred from the third-person perspective. This is problematic, too, because we do not experience mental states in a third-person perspective. Instead, we observe physical states in the third-person perspective. One may consequently infer mental brain properties from the observation of physical states. This inference also remains questionable, since physical states can be entirely accounted for by physical properties without need to infer any mental brain properties. Accordingly, either type of inference of mental brain properties from physical brain properties lacks evidence.

The criterion of physical–mental intimacy may be considered in a novel light in the context of the postulated world–brain relation. If the world-brain relation is a necessary condition of possible mental states, one may replace physical–mental intimacy by 'neuro-ecological intimacy'. 'Neuro-ecological intimacy' refers to the fact that the brain stands in an intimate and mutually dependent relationship with its respective ecological context. Only if there is such 'neuro-ecological intimacy' can we have mental states at all. Loss of mental states, as in loss of consciousness, goes along with loss of world–brain relation and its neuro-ecological intimacy. For instance, patients in the vegetative state have lost consciousness and are no longer able to relate to the world in a meaningful way—their brains have lost their world–brain relation and are henceforth no longer relational but isolated (see what follows for more details).

Organizational and Functional Dissociation

An interesting feature in McMahan's account is the possible dissociation between organizational and functional brain continuity. McMahan's insistence on minimal functional continuity as the bearer of personal identity—without access to the particular content of mental states such as memories—suggests that what he has in mind is phenomenal continuity. Since McMahan thinks that this sort of phenomenal continuity is enough to secure the distinct concern for one's own future, it remains to be shown how minimal functional continuity can do the trick. In situations of phenomenal disruptions of consciousness (such as falling into a dreamless sleep), one wonders how one could tell that there is any connection between a conscious state before and after the disruption occurs. If there is no psychological continuity whatsoever, it seems that there is in principle no way to differentiate between one's own conscious state and someone else's conscious state just by reference to the first-person perspective. The only way to be sure of one's own identity, following McMahan, would then be to see a neurologist and have her check if it is still the same minimal functional continuity that supports my phenomenally disconnected states of consciousness. This, of course, would reduce personal identity to the third-person perspective and thus seems inadequate as a basis for egoistic concern about the future.

22.1 Spotlight: Mind–Body Identity: Are We Just Our Brains?

Kimberly Van Orman

Before we can consider questions about personal identity, free will, consciousness and others in the philosophy of mind, we need an answer to the metaphysical question: What is the relationship between the mind and the body? There are many possibilities available in the logical space: that only nonphysical minds exist (idealism), only brains are real (eliminative materialism), minds and brains both exist but are essentially the same thing (reductive materialism), minds and brains exist but are fundamentally different things (substance dualism), minds are a nonphysical property of brains (property dualism) and minds are not nonphysical but are defined in terms of their functional properties, like software that runs on brains (nonreductive materialism). Of these, four come up most frequently: two dualist and two materialist views.

Dualism is committed to the idea that minds are fundamentally distinct from brains. There are two main forms of dualism: substance and property. *Substance dualism* is most

associated with René Descartes, who held that minds are made of a nonphysical substance distinct from physical brains and bodies, although the two different substances can interact. This is a view that easily and intuitively fits with the idea of a soul that can survive the death of the body. Few philosophers today accept this view as likely to be true. The centrality of the brain to mind is evident in cases where physical damage to the brain seems to result in a person changing in ways fundamental to who they are as a person. That should not happen if substance dualism is true. In addition, if the mind were a completely nonphysical substance, we would need an account of how something with absolutely no physical properties could have any effect on a world in which physical events seem to always have physical causes.

Property dualism doesn't require that the mind be a completely separate substance. It accepts the brain as the seat of the mind but holds that the mind is an emergent property that cannot be reduced to the physical. Property dualism can better account for what we know about the brain; however, it still has a problem explaining how the nonphysical mind can make physical changes in the brain. Some philosophers committed to property dualism have given up on the idea that minds can have causal properties. On this view, called *epiphenomenalism*, minds are both nonphysical and noncausal properties that arise out of the functioning of the brain.

With regard to neuroethical concerns about identity and authenticity that arise, for instance, in the context of cognitive enhancement, substance dualism would deny that we, as distinct individuals, are essentially our brains. This implies that enhancements that specifically target physical brain function would not threaten an identity whose locus is the nonphysical mind. For property dualism, it's possible that changes to the brain could affect one's mind.

Every version of materialism (also known as physicalism) is committed to the idea that there is only one basic kind of matter in the world—physical matter. The principle of *supervenience* describes a dependency relation that attempts to capture this notion that everything is in some sense physical. When an entity or property (e.g., a mind) supervenes on another (e.g., a brain), then if they share all the same physical properties (down to the subatomic level), they share the same mental ones.

Reductive materialism is the idea that the mind and the brain are just two ways of talking about the same thing. On this view, for any type of mental state (for example, being in a specific kind of pain), there will be an underlying physical type. In theory, we could give a physical definition for any mental state (since they're the same thing). For example, pain of type *p* would be defined in terms of the physical brain states that it's identical with. This view is also called *type identity theory*. This take on materialism captures the intuitive idea that minds and brains are simply the same thing, just described in different language. Changes to the brain also constitute changes to the mind. To the extent that we are concerned about the integrity, authenticity or identity of the self, reductive materialism holds that brain and mind are identical, which implies that changes to the brain would constitute changes to the mind/self.

Nonreductive materialism accepts supervenience and holds that minds are fundamentally physical. However, it does not accept the idea that minds and brains are identical. A mental state is defined as a specific kind (for example, a pain, an emotion, an intention) based on its relationships to other mental states, the situation of the body (senses, physical location, condition) and the behaviors it gives rise to. For a nonreductive materialist, minds are to

brains as software is to hardware. There is nothing nonphysical going on, but at least some of the causal force comes from what's going on in the mind. This view is sometimes called *token identical materialism* since it holds that any individual (token) mental state will be identical with a token brain state. It denies that we can get type identity because it holds that some mental states can be multiply realizable—for example, my mental state of affection for my cat might be realized in my brain differently than your (identical) mental state is realized in yours. However, even though our particular brain structure might differ, we share the same type of mental state, since it comes up when we each see our cats and leads us to walk over and scratch their ears and leads to other cat-friendly thoughts. For nonreductive materialism (as in property dualism), whether targeted or widespread alterations in the function of the physical brain would threaten personal identity will depend on the nature of the alterations and whether we see a significant change in mental status. Minds and brains are not identical, but they are interdependent. Some problems will be related to the software (mind), and others will be related to the hardware (brain). Given some configurations of the hardware, the software won't function properly. Similarly, for some changes in the brain (such as we see with severe trauma), mental states could be changed enough to raise concerns about whether one's identity has changed.

~~Bibliography~~

Chalmers, D. (2002) *Philosophy of Mind: Classical and Contemporary Readings*. Oxford University Press.

Montero, B. (2008) *On the Philosophy of Mind*. Wadsworth Publishing.

Ravenscroft, I. (2005) *Philosophy of Mind: A Beginner's Guide*. Oxford University Press.

'World–Brain Relation' and First-/Third-Person Perspective

One may now want to raise the question of how our criterion of world–brain relation stands in relationship to the distinction between first- and third-person perspective. The discussion of personal identity revolves largely around the question of whether the first- and/or the third-person perspective provides necessary and/or sufficient conditions for personal identity. We argue that the world–brain relation cannot be accessed through the third-person perspective since it cannot be observed as such. Nor can the world–brain relation be experienced as such via the first-person perspective since it is prephenomenal rather than phenomenal (Northoff, 2014b). How then can we characterize the world–brain relation in perspectival terms? We assume that the world–brain relation remains by itself nonperspectival, meaning that it is neither first- nor third-personal.

How, though, can we then apprehend and access the world–brain relation if not through the first- and third-person perspective? We claim that we can access the world–brain relation only in an indirect way by inferring from combined experience and observation in very much the same way that Nagel suggests the need for a correspondence between the first- and third-person perspectives. At the same time, a proper world–brain relation may be conceived of as a necessary or predisposing condition for the possible perspectival differentiation between the first- and third-person perspectives (Northoff, 2014b). One may consequently feature the world–brain relation as preperspectival rather than as completely nonperspectival or perspectival: the prefix

'pre–' indicates here that the world–brain relation is the necessary condition of subsequent perspectival access to reality in terms of first-, second- and/or third-person perspectives. In contrast, it would be nonperspectival when there is no relationship at all between world–brain relation and the different perspectives. Why is the distinction between pre- and nonperspectival important for personal identity? We postulate that such preperspectival features need to be continuous in order to preserve our personal identity. If the world–brain relation is disrupted, becoming nonperspectival rather than preperspectival, we may lose our personal identity, including our relation to the world. For instance, patients in vegetative state, who have lost consciousness, lose the neuro-ecological intimacy of their world–brain relationship, which renders the latter nonperspectival. That, in turn, makes it impossible for them to take on any kind of perspective, neither first- nor second-person perspective (as manifest in the absence of first-person experience and second-person intersubjectivity) nor third-person perspective (as manifest in the absence of third-person observation).

Neuroethical Considerations and Future Directions

We now turn to briefly indicate how some of the discussed features relate to neuroethical concerns of personal identity. We do so by looking at patients in the vegetative state and at dementias. Questions about the personal identity of these patients might affect decision-making authority over these patients, including the authority of previously expressed desires in advance directives. We will indicate how a world–brain relation and neuro-ecological account of personal identity illuminate such hard cases.

Patients in a permanent vegetative state have irreversibly lost consciousness. These patients are unable to relate to the world in a meaningful way. In our terminology, their world–brain relation is disrupted, rendering their personal identity discontinued. Even though these patients are biologically alive, in that their vital functions are artificially sustained, there's nothing left of the relational component that would enable the person that previously inhabited the biological organism to relate to the world. Due to this lack of preperspectival world–brain relation, there is no more first-personal experience, and so one might conclude that the person has vanished. What is left is a 'nonrelational organism' on life support—an organism that once was a constitutive part of a person's identity. This description might paint a picture of how to ontologically characterize these patients, but does it allow for a clear-cut neuroethical evaluation of these scenarios when it comes to previously expressed desires in advanced directives? This issue is far from being straightforward and comes with a substantial commitment to normative ethical convictions. We won't attempt a decisive answer here but merely hint at two possible interpretations. On one view, the person that previously expressed their desire to, say, be kept on life support even if their consciousness is irreversibly lost might be seen to have normative authority over an organism that once was the host of their personhood, even if the person vanished. This is so, perhaps, because there is no other person inhabiting that organism who could claim authority. On a different interpretation pertaining to the same point, one might aver that precisely because there is no person inhabiting that organism anymore, a previously expressed directive has no authority, and so it should be up to the closest relatives or guardian to decide.

Cases of dementia, especially in its very late stages, are even more difficult to evaluate both ontologically and normatively. There is no question that in early stages of dementia, the person still relates to the environment in their usual way. When the gradual decline of person-characteristics proceeds, the way in which the person relates to their environment potentially changes in a great many different ways. Nonetheless, there seems to be still enough of the world–brain relation in place so as to preserve personal identity, not necessarily because of the

contents of psychological continuity that are inevitably lost toward the end of that neurodegenerative disease but because there is a continuous uninterrupted way in which the same person relates to their environment due to an ongoing world–brain relation. The person might have lost some of their distinct psychological features, but the very basis for their being able to relate to the world remains, and so their personal identity is preserved. If this is so, one would think, there are at least prima facie reasons to make the case that previously expressed desires should continue to bear normative authority even if the person in their current stage can no longer remember having expressed these desires. It goes without saying that this issue becomes ever more difficult if current desires contradict previously expressed ones.

Both scenarios illustrate how a relational account of personal identity might be able to offer empirically informed normative guidance in cases where purely metaphysical views of personal identity either decide by fiat or leave the issue open entirely. Needless to say, a relational look at these cases is still in the fledgling stage, and so it becomes apparent that future interdisciplinary neuroethical studies are needed in order to get a more firm grip on these vexing issues.

In conclusion, future investigations of personal identity may want to discuss (1) the model of the brain; (2) the brain's relationship to the world; (3) the interplay between world–brain relation and consciousness, including its epistemological features like auto-epistemic limitation; and (4) the notion of brain continuity as in regards to both world–brain relation and consciousness. This may, in turn, provide the ground for the future development of a brain-based and neuro-ecological (rather than brain-reductive and neuronal) account of personal identity.

Further Reading

DeGrazia, D. (1999) "Persons, Organisms, and the Definition of Death: A Philosophical Critique of the Higher-Brain Approach". *Southern Journal of Philosophy* 37, 419–440.

Lewis, D. (1976) "Survival and identity". In A. Rorty (Ed.). *The Identities of Persons*. Berkeley, CA: University of California Press, pp. 17–41.

MacIntyre, A. (1984) *After Virtue*. Notre Dame: University of Notre Dame Press.

Noonan, H. (2003) *Personal Identity*. 2nd ed. London: Routledge.

Northoff, G. (2014a) *Unlocking the Brain. Volume I: Coding*. New York: Oxford University Press.

Northoff, G. (2014b) *Unlocking the Brain. Volume II: Consciousness*. New York: Oxford University Press.

Schechtman, M. (2014) *Staying Alive—Personal Identity, Practical Concerns, and the Unity of a Life*. New York: Oxford University Press. doi:10.1093/acprof:oso/9780199684878.001.0001

References

Atteveldt van, N., *et al.* (2014) Multisensory integration: Flexible use of general operations. *Neuron* 81(6): pp. 1240–1253.

Duncan, N.W., Hayes, D.J., Wiebking, C., Tiret, B., Pietruska, K., Chen, D.Q., . . . Northoff, G. (2015) Negative childhood experiences alter a prefrontal-insular-motor cortical network in healthy adults: A preliminary multimodal rsfMRI-fMRI-MRS-dMRI Study. *Human Brain Mapping* 36(11): pp. 4622–4637.

Evans, G.M. (1985) *Collected Papers*, Antonia Phillips (Ed.). Oxford: Clarendon.

Kuhlenbeck, H. (1965) The concept of consciousness in neurological epistemology in brain and mind. In J.R. Smythies (Ed.). *Brain and Mind: Modern Concepts of the Nature of the Mind*. New York: Routledge, pp. 102–143.

Lewis, D. (1976) Survival and identity. In A.O. Rorty (Ed.). *The Identities of Persons*. Berkeley: University of California Press.

Locke, J. (1694/1975) *An Essay Concerning Human Understanding,* P. Nidditch (Ed.). Oxford: Clarendon Press.

McMahan, J. (2002) *The Ethics of Killing: Problems at the Margins of Life*. Oxford: Oxford University Press.
Nagel, T. (1971) Brain bisection and the unity of consciousness. *Synthese* 22: pp. 396–413.
———. (1979) *Mortal Questions*. London: Routledge.
———. (1986) *The View from Nowhere*. Oxford: Oxford University Press.
———. (1987) *What Does It All Mean? A Very Short Introduction to Philosophy*. New York, Oxford: Oxford University Press.
Northoff, G. (2001) *Personale Identität und Operative Eingriffe in das Gehirn*. Paderborn: Mentis.
———. (2004a) Am I my brain? Personal identity and brain identity—A combined philosophical and psychological investigation in brain implants. *Philosophia Naturalis* 41: pp. 257–282.
———. (2004b) *Philosophy of the Brain: The Brain Problem*. Amsterdam: John Benjamins.
———. (2014) *Unlocking the Brain. Vol. II: Consciousness*. New York, Oxford: Oxford University Press.
———. (2016) *Neurophilosophy and the Healthy Mind. Learning from the Unwell Brain*. New York: Norton.
Olson, E. (1997) *The Human Animal: Personal Identity Without Psychology*. New York: Oxford University Press.
Parfit, D. (1984) *Reasons and Persons*. Oxford: Clarendon Press.
Shoemaker, S. (1963) *Self-knowledge and Self-identity*. Ithaca, NY: Cornell University Press.
Snowdon, P. (2014) *Persons, Animals, Ourselves*. New York: Oxford University Press.
Sperry, R. (1966) Brain bisection and mechanisms of consciousness. In J.C. Eccles (Ed.). *Brain and Conscious Experience*. New York: Springer, pp. 84–108.
Weinberg, S. (2011) Locke on personal identity. *Philosophy Compass* 6(6): pp. 398–407.
Whitehead, A.N. (1929/1978) *Process and Reality*. New York: Free Press.
Wiggins, D. (1967) *Identity and Spatio-Temporal Continuity*. Oxford: Wiley-Blackwell.
Wilkes, K. (1988) *Real People*. Oxford: Clarendon Press.
Williams, B. (1970) The self and the future. *Philosophical Review* 79(2): pp. 161–180.